Robust Inference for Projection Structures

Nurushev, N.

2019

document version
Publisher's PDF, also known as Version of record

ROBUST INFERENCE FOR PROJECTION STRUCTURES

VRIJE UNIVERSITEIT

ROBUST INFERENCE FOR PROJECTION STRUCTURES

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op maandag 11 februari 2019 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Nurzhan Muratovich Nurushev

geboren te Aktobe, Kazachstan

promotor: dr. E.N. Belitser
copromotor: dr. P. J. de Andrade Serra

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my supervisor Eduard Belitser for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my PhD study.

I am also indebted to the members of my defense committee: Subhashis Ghosal, Botond Szabó, Aad van der Vaart, Mark van de Wiel, Wessel van Wieringen and Harry van Zanten. Thank you for your willingness to be in the committee, the time you spent reading this thesis and for your helpful feedback. I am also thankful to my co-promoter Paulo Serra for helpful discussions.

Furthermore, I would like to thank my paranymphs, Alex and Ivan. I am also grateful to my colleagues of VU Amsterdam: Mathisca, Marielle, Fetsje, Bartek, Maarten, Beata, Mehran, Gwen, Birgit, Tim, Dennis, Rikkert, Maria, Marina an Luminita. Thank you all for the lunch and coffee breaks, and various celebrations. Special thanks to the members of the Bayes club for interesting talks and discussions.

I have made so many friends since I came to Amsterdam in February 2014. Some of them are already far away from me, but I remember how wonderful we spent our time together. I would especially like to mention my close buddies: Arai, Yerlan, Madina and Zarina. Thank you for everything, you know how to cheer me up when I am not in the mood.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and my life in general. I am also grateful to my sisters, Gaukhar and Assel, for taking care of our parents during my absence. These things are never overrated, especially when you are thousands of kilometers apart. Without my family, this thesis would really never exist.

SUMMARY

Nonparametric and *high-dimensional models*, due to their high flexibility, are widely used in statistics to find a better approximation for the underlying mechanism generating the observed data. There exists a huge number of statistical methods to make inference on these models.

In this thesis we focus on four important statistical inference problems: estimation, posterior contraction, structure recovery (in a weak sense) and uncertainty quantification, by using empirical Bayes and penalization methods. The main contribution of this thesis is development of a general robust framework for addressing the above mentioned inference problems and applying these to a number of various examples of high-dimensional and nonparametric models and structures, under possible misspecification.

In Chapter 1, we give a brief introduction to some important theoretical notions and concepts which we are going to use in the sequel.

In Chapter 2, we obtain novel results for the grand problem of uncertainty quantification (on the way solving other inference problems as well) for possibly sparse sequences.

In Chapter 3, we consider empirical Bayesian inference in the many normal means model in the situation when the high-dimensional mean vector is *multilevel sparse*.

In Chapter 4, by using the penalization method, we study the following inference problems in the *biclustering model*: estimation, structure recovery (in a weak sense) and *uncertainty quantification*.

Finally, in Chapter 5, we develop the *general framework of projection structures* and study the above mentioned inference problems within this framework by using *empirical Bayes* and *penalization* methods.

This thesis is mainly based on the following publications:

BELITSER, E. and NURUSHEV, N. (2018). Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences. Under revision for *Bernoulli* (see also ArXiv:1511.01803).

BELITSER, E. and NURUSHEV, N. (2017). Local posterior concentration rate for multilevel sparse sequences. *Bayesian Statistics in Action*, Springer Proc. Math. Stat. 194, 51–66.

BELITSER, E. and NURUSHEV, N. (2018). Local inference by penalization method for biclustering model. *Math. Methods Statist.* 27, 163–183.

BELITSER, E. and NURUSHEV, N. (2018). Robust inference for general projection structures by empirical Bayes and penalization methods. *Submitted*.

SAMENVATTING

Niet-parametrische en *hoog-dimensionale modellen* worden, vanwege hun hoge flexibiliteit, veel gebruikt in de statistiek om de onderliggende data-genererende mechanisme te vinden voor de waargenomen gegevens. Er bestaat een groot aantal statistische methoden om conclusies te trekken over deze modellen.

In dit proefschrift concentreren we ons op vier belangrijke statistische problemen: *schatting, contractie van de a-posteriori verdeling, structuurherstel* (in een zwakke zin) en *onzekerheid-kwantificering*, door gebruik te maken van *empirische Bayes* en *penalisatie* methoden. De belangrijkste bijdrage van dit proefschrift is de ontwikkeling van een *algemeen robuust raamwerk* voor het aanpakken van de bovengenoemde problemen en dit toe te passen op een aantal voorbeelden van hoog-dimensionale en niet-parametrische modellen en structuren onder mogelijke misspecificatie.

In Hoofdstuk 1 introduceren we enkele belangrijke theoretische begrippen en concepten die we in het vervolg gebruiken.

In Hoofdstuk 2 leiden we nieuwe resultaten af voor het hoofdprobleem van de onzekerheid-kwantificering (onderweg lossen wij andere inferentie-problemen op) voor mogelijk spaarzame vector waarvan veel (vooraf onbekende) coördinaten nullen zijn.

In Hoofdstuk 3 bestuderen we het schattingsprobleem in het normale "signaal+ruis" model gebruikmakend van de empirische Bayesiaanse aanpak, in de situatie dat de hoog-dimensionale signaal *multi-niveau spaarzaam* is.

In Hoofdstuk 4 bestuderen we de volgende inferentie-problemen in het *biclustering model*: schatting, structuurherstel (in de zwakke zin) en de onzekerheid-kwantificering, door gebruik te maken van de penalisatie methode.

Ten slotte ontwikkelen we in Hoofdstuk 5 het *algemene raamwerk van projectie-structuren* en bestuderen we de bovengenoemde inferentie-problemen in dit algemene kader, gebruikmakend van de empirische Bayes en penalisatie methoden.

Dit proefschrift is grotendeels gebaseerd op de volgende publicaties:

BELITSER, E. and NURUSHEV, N. (2015). Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences. Under revision for *Bernoulli* (see also ArXiv:1511.01803).

BELITSER, E. and NURUSHEV, N. (2017). Local posterior concentration rate for multilevel sparse sequences. *Bayesian Statistics in Action*, Springer Proc. Math. Stat. 194, 51–66.

BELITSER, E. and NURUSHEV, N. (2018). Local inference by penalization method for biclustering model. *Math. Methods Statist.* 27, 163–183.

BELITSER, E. and NURUSHEV, N. (2018). Robust inference for general projection structures by empirical Bayes and penalization methods. *Submitted.*

CONTENTS

NOTATION

$\mathbb{R}, \mathbb{R}_+$	real numbers, non-negative real numbers
$\mathbb{N}$	natural numbers
$\mathbb{N}_0$	$\{0\} \cup \mathbb{N}$
$\mathbb{R}^n$	n-dimensional Euclidian space
$[k] = \mathbb{N}_k$	$\{1, \dots, k\}$
$[k]_0$	$0 \cup [k],\ k \in \mathbb{N}$
$\lesssim, \gtrsim$	less or equal (resp. larger or equal) up to a universal constant
$\ll, \asymp$	of a smaller order, of the same order
$1_E = 1\{E\}$	indicator function of the event E
I	identity matrix
$\lvert A \rvert$	cardinality of the set A
$A \setminus A_0$	$\{s \in A : s \notin A_0\}$
A^c	complement of A
$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{N}(\mu, \Sigma)$	multivariate normal distr. with mean μ and covariance matrix Σ
$\phi(x, \mu, \sigma^2)$	density of $\mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$ at point x, where $Z \sim \mathrm{N}(0,1)$
$\varphi(x, \mu, \Sigma)$	density of $\mathrm{N}(\mu, \Sigma)$ at point x
$\mathrm{N}(\mu, 0) = \delta_\mu$	Dirac measure at point μ
$a \vee b$	maximum between a and b
$a \wedge b$	minimum between a and b
$\lfloor a \rfloor$	$\max\{m \in \mathbb{Z} : m \le a\}$
$\triangleq$	equality by definition
$\langle x, y \rangle$	usual scalar product between $x, y \in \mathbb{R}^n$
$\mathrm{P}_{\mathbb{L}_I} = \mathrm{P}_I$	projection operator onto the linear subspace $\mathbb{L}_I$
$\mathrm{P}_I^\perp = \mathrm{I} - \mathrm{P}_I$	projection operator onto the orthogonal complement $\mathbb{L}_I^\perp$ of $\mathbb{L}_I$

1

INTRODUCTION

In this chapter we give a brief introduction to some important theoretical notions and concepts which we are going to use in the sequel, such as *nonparametric* and *high-dimensional* statistical models, global (*minimax*) and local (*oracle*) rates, *adaptation* problem, relation between oracle and minimax results, traditional *oracle inequalities* in expectation and non-asymptotic local exponential deviation bounds in probability, *(empirical) Bayesian approach*, global and local *posterior contraction rates*, *uncertainty quantification problem*, *deceptiveness* issue in uncertainty quantification, *structures*, *structural layers*, *layer complexity*, *robust* inference, *exchangeable exponential moment condition*. Also we describe three main problems studied in this thesis: *estimation*, *posterior contraction* and *uncertainty quantification*. Then we introduce a general framework of projection structures under *misspecification*, for which we study the above mentioned inference problems. This chapter is concluded with an overview section.

1.1. STATISTICAL MODELING

The central object of data analysis is the model and its relation with the observed data. In this thesis we observe a random element Y coming from some unknown probability distribution $\mathbb{P}_0^{(\epsilon)}$, which belongs to some model $\mathcal{P}^{(\epsilon)}$, on a measurable space $(\mathcal{Y}^{(\epsilon)}, \mathcal{F}^{(\epsilon)})$, where $\mathcal{P}^{(\epsilon)}$ is a collection of probability distributions, $\mathcal{Y}^{(\epsilon)}$ is some metric space and $\mathcal{F}^{(\epsilon)}$ is a Borel σ-algebra of $\mathcal{Y}^{(\epsilon)}$. The parameter ϵ is known and plays in some sense a role of the "amount of information" in the model. It can be, for instance, $\epsilon = \sigma$ or $\epsilon = N^{-1/2}$, where σ is the variance of an additive noise and N is the number of observations in the sample. In asymptotic regimes one usually considers: decreasing noise level $\sigma \to 0$, high-dimensional setup $N \to \infty$, or their combination, e.g., $\sigma = N^{-1/2}$ and $N \to \infty$. Thus, the behavior of the statistical procedure can be qualified in terms of parameter ϵ. Often we drop the dependence on ϵ, e.g., $\mathcal{Y} = \mathcal{Y}^{(\epsilon)}, Y = Y^{(\epsilon)}$, unless we want to emphasize this dependence for some reason.

Usually the model $\mathcal{P}$ is characterized by a parameter θ belonging to a set Θ:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

1

If $\Theta \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, then the model $\mathcal{P}$ is said to be *parametric*. If Θ is an infinite dimensional space (say, functions or sequences), then we call $\mathcal{P}$ a *nonparametric* model. One of the new paradigms in statistical theory, which has been developed in the last decades, is *high-dimensional* data (model). Informally, if $\mathcal{Y} = \mathbb{R}^N$ and $\Theta = \mathbb{R}^d$, the model $\mathcal{P}$ is said to be *high-dimensional*, if the parameter dimension d is of the same order or much larger than the sample size N. The main statistical problem is to make an inference on the parameter θ. For instance, construct an estimator and a confidence set around it, describing the uncertainty of the estimator. In high-dimensional models, "good" inference is in general impossible unless θ has a *structure*; for example, the majority of the coordinates of high-dimensional $\theta \in \mathbb{R}^d$ are zeros whereas the locations of few nonzero coordinates are not known to the observer.

In the following sections, among other things, we briefly describe two approaches to statistical inference: *frequentist* and *Bayesian*.

1.2. FREQUENTIST INFERENCE: MINIMAX VERSUS ORACLE

One of the main problems in statistical theory is to construct an estimator for the underlying θ by using the observed data Y. An *estimator* is a measurable function of the data $\hat{\theta} = \hat{\theta}(Y) : \mathcal{Y} \mapsto \Theta$. Since the estimator $\hat{\theta} = \hat{\theta}(Y)$ varies along the data Y, one of the ways to measure the closeness of the estimator $\hat{\theta}$ to the unknown parameter θ is in average. Consider a loss function $L : \Theta \times \Theta \mapsto \mathbb{R}_+ = [0, +\infty]$, which quantifies how the estimator $\hat{\theta}$ deviates from the true parameter θ. It is common to consider the loss functions L as some metric d on Θ, or its power. Then we can measure the quality of the estimator $\hat{\theta}$ by the risk function $R^2(\hat{\theta}, \theta) = \mathbb{E}_\theta L(\hat{\theta}, \theta)$, where $\mathbb{E}_\theta$ denotes the expectation with respect to the measure $\mathbb{P}_\theta$.

Suppose that the true parameter $\theta \in \Theta_s \subseteq \Theta$, indexed by the so called *structural parameter* $s \in \mathcal{S}$ (e.g., smoothness or sparsity) which is assumed to be known. Then one of the benchmarks in the estimation problem is the *minimax rate* (or *minimax risk*)

$$r^2(\Theta_s) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta_s} R^2(\hat{\theta}, \theta), \tag{1.1}$$

where the infimum is taken over all possible estimators $\hat{\theta} = \hat{\theta}(Y)$. The minimax rate provides the least possible mean loss when the worst case happens. In a way, it expresses the complexity of the estimation problem over the set Θ_s. The goal in this problem is to find a so called minimax estimator $\hat{\theta}$, which attains the minimax rate in the sense that

$$\sup_{\theta \in \Theta_s} R^2(\hat{\theta}, \theta) \leq C r^2(\Theta_s),$$

for some absolute $C \geq 1$. This inequality becomes stronger as C gets closer to 1.

So far we could use knowledge of structural parameter s. However, in practice the structural parameter is often not known in advance. It is therefore desirable to find another procedure that can achieve the minimax rate without knowledge of $s \in \mathcal{S}$. That is, we are given a scale $\{\Theta_s, s \in \mathcal{S}\}$ and we only know that the parameter $\theta \in \cup_{s \in \mathcal{S}} \Theta_s \subseteq \Theta$. This is the *adaptation* problem. The goal of this problem is to construct an adaptive minimax

estimator $\hat{\theta} = \hat{\theta}(Y)$ (i.e., without knowledge of structural parameter s) such that

$$\sup_{\theta \in \Theta_s} R^2(\hat{\theta}, \theta) \lesssim r^2(\Theta_s), \tag{1.2}$$

for all $s \in \mathcal{S}$. Several adaptive estimation methods are developed in the frequentist statistics: blockwise method ([39], [34]), Lepski's method ([58], [59]), wavelet thresholding ([37]), penalization method (see [8], [23] and further references therein). Some of these methods are applicable only for specific settings: for instance, blockwise method is originally designed for the white noise sequence model with the mean squared risk. Some methods are more general, for example, Lepski's and penalization methods can be extended to different settings (various risk function).

Besides the minimax (global) approach, there is a way to address the adaptation problem from another perspective. Namely, a so called *oracle* (local) approach turns out to be more flexible than the minimax approach. Assume that we have a family of estimators $\hat{\Theta} = \hat{\Theta}(\mathcal{I}) = \{\hat{\theta}(I), I \in \mathcal{I}\}$, where $\mathcal{I}$ is a collection of the so called *structures* I such that $\inf_{I \in \mathcal{I}} R^2(\hat{\theta}(I), \theta) = R^2(\hat{\theta}(I_o))$ is attained at some $I_o = I_o(\theta) \in \mathcal{I}$ for any $\theta \in \Theta$. This means that, for each $\theta \in \Theta$, $I_o = I_o(\theta)$ is the best structure at which we achieve the smallest possible rate (called *oracle rate*)

$$r^2(\theta) = r^2(\hat{\Theta}, \theta) = \inf_{\hat{\theta} \in \hat{\Theta}} R^2(\hat{\theta}, \theta) = \inf_{I \in \mathcal{I}} R^2(\hat{\theta}(I), \theta) = R^2(\hat{\theta}(I_o), \theta), \tag{1.3}$$

where $\hat{\theta}(I_o)$ is called the *oracle* estimator.

Thus, the oracle rate is our new benchmark in this approach and the main goal is to find an estimator $\hat{\theta} = \hat{\theta}(\hat{I})$ with some $\hat{I} = \hat{I}(Y)$ such that for any $\theta \in \Theta$ and for some constants $C_0, C_1 > 0$ the following *oracle inequality* is fulfilled:

$$R^2(\hat{\theta}(\hat{I}), \theta) \le C_0 r^2(\theta) + C_1 p(\epsilon), \tag{1.4}$$

where $p(\epsilon) \ge 0$ is some penalty term and the oracle rate $r^2(\theta)$ is defined by (1.3). Typically $p(\epsilon) \lesssim r^2(\theta)$ (often even $p(\epsilon) \ll r^2(\theta)$ as $\epsilon \to 0$). The inequality (1.4) becomes stronger as C_0 gets closer to 1, C_1 gets smaller and the space Θ gets bigger. Another, stronger form of oracle inequality is the exponential non-asymptotic bound in probability for $L(\hat{\theta}(\hat{I}), \theta)$: there exist positive constants M_1, H_1, m_1 such that for any $\theta \in \Theta$ and any $M \ge 0$,

$$\mathbb{P}_\theta\big(L(\hat{\theta}(\hat{I}), \theta) \ge M_1 r^2(\theta) + M p(\epsilon)\big) \le H_1 e^{-m_1 M}. \tag{1.5}$$

Clearly, (1.5) is stronger than (1.4). Indeed, denoting $U = L(\hat{\theta}(\hat{I}), \theta)$, $v = u - M_1 r^2(\theta)$, from (1.5) it follows that

$$\begin{aligned}
\mathbb{E}_\theta L(\hat{\theta}(\hat{I}), \theta) = \mathbb{E}_\theta U &= \int_0^\infty \mathbb{P}_\theta(U > u)\, du \le M_1 r^2(\theta) + \int_{M_1 r^2(\theta)}^\infty \mathbb{P}_\theta(U > u)\, du \\
&= M_1 r^2(\theta) + \int_0^\infty \mathbb{P}_\theta(U > M_1 r^2(\theta) + v)\, dv \le M_1 r^2(\theta) + H_1 \int_0^\infty e^{-\frac{m_1 v}{p(\epsilon)}}\, dv \\
&= M_1 r^2(\theta) + \frac{H_1 p(\epsilon)}{m_1}.
\end{aligned}$$

Notice that if $p(\epsilon) \lesssim r^2(\theta)$, then $\mathbb{E}_\theta L(\hat\theta(\hat I),\theta) \lesssim r^2(\theta)$. Moreover, the non-asymptotic exponential probabilistic bound (1.5) is finer than the traditional oracle inequality (1.4) in expectation. This refined formulation allows to consider various asymptotic regimes ($N \to \infty$, $\epsilon \to 0$, or their combination) as we can let M depend in any way on N, ϵ, or both. In this thesis we are mostly interested in non-asymptotic exponential bounds in probability of type (1.5), when solving the estimation problem.

A natural question arises whether the two adaptation optimality procedures, minimax over the scale $\{\Theta_s, s \in \mathcal{S}\}$ and oracle over the family of estimators $\hat\Theta = \hat\Theta(\mathcal{I}) = \{\hat\theta(I), I \in \mathcal{I}\}$, are related to each other. Suppose that for any $s \in \mathcal{S}$ there exists $I = I(s) \in \mathcal{I}$ such that $\hat\theta(I(s))$ is minimax over the class Θ_s:

$$\sup_{\theta\in\Theta_s} R^2(\hat\theta(I(s)),\theta) \le C' r^2(\Theta_s), \tag{1.6}$$

ideally with $C' = 1 + o(1)$ as $\epsilon \to 0$, otherwise with some uniform constant $1 \le C'' < \infty$. Then we say that the family of estimators $\hat\Theta(\mathcal{I})$ *covers* the scale $\{\Theta_s, s \in \mathcal{S}\}$.

Suppose the oracle inequality (1.4) holds for a family of estimators $\hat\Theta(\mathcal{I})$ which covers a scale $\{\Theta_s, s \in \mathcal{S}\}$, and suppose further that $p(\epsilon) \le c r^2(\Theta_s)$. Then we claim that the oracle inequality (1.4) implies the adaptive minimax result over the scale $\{\Theta_s, s \in \mathcal{S}\}$, i.e., the estimator $\hat\theta(\hat I)$ from (1.4) is adaptive minimax with respect to the scale $\{\Theta_s, s \in \mathcal{S}\}$. Indeed, in view of (1.4), (1.6) and $p(\epsilon) \le c r^2(\Theta_s)$, we obtain that for all $s \in \mathcal{S}$,

$$\begin{aligned}\sup_{\theta\in\Theta_s} R^2(\hat\theta(\hat I),\theta) &\le C_0 \sup_{\theta\in\Theta_s} \inf_{I\in\mathcal{I}} R^2(\hat\theta(I),\theta) + C_1 p(\epsilon) \le C_0 \sup_{\theta\in\Theta_s} R^2(\hat\theta(I(s)),\theta) + C_1 p(\epsilon)\\ &\le (C_0 C' + c C_1) r^2(\Theta_s).\end{aligned} \tag{1.7}$$

This means that the local (oracle) approach for an appropriately chosen family of estimators $\hat\Theta(\mathcal{I})$ (which covers the scale $\{\Theta_s, s \in \mathcal{S}\}$) is stronger than the global (adaptive minimax) approach.

From (1.6), it follows that $r^2(\theta) \le C' r^2(\Theta_s)$ for all $\theta \in \Theta_s$, $s \in \mathcal{S}$. This and (1.5) imply the following adaptive minimax result: for any $M \ge 0$

$$\sup_{\theta\in\Theta_s} \mathbb{P}_\theta\big(L(\hat\theta(\hat I),\theta) \ge M_1 C' r^2(\Theta_s) + M p(\epsilon)\big) \le H_1 e^{-m_1 M}. \tag{1.8}$$

Thus, the oracle result (1.4) (or (1.5)) implies the adaptive minimax result (1.2) (or (1.8)) over all scales which are contained in the space Θ and covered by the family of estimators $\hat\Theta(\mathcal{I})$, i.e., for which (1.6) is fulfilled. Notice that this claim holds only if the family of estimators $\hat\Theta(\mathcal{I})$ is appropriately chosen: the family $\hat\Theta(\mathcal{I})$ should neither be too poor, nor too rich. Indeed, if the family is too rich, then it may not be possible to find an estimator $\hat\theta(\hat I)$ such that (1.4) (or (1.5)) is fulfilled for a reasonable Θ. Instead, (1.4) (or (1.5)) may hold only for a "thin" space Θ, but we would like Θ to be as big as possible, which contains all $\Theta_s, s \in \mathcal{S}$. An oracle approach to optimality of estimators was probably first studied by [56], within the class of ordered linear smoothers and then later developed in the series of works by Donoho and Johnstone.

To illustrate the introduced concepts, consider the following example.

Example: signal+noise model. Assume we observe the data

$$Y_i = \theta_i + \tfrac{1}{\sqrt{n}}\xi_i, \quad i \in \mathbb{N},$$

where $\theta = (\theta_i)_{i\in\mathbb{N}} \in \Theta = \ell_2$ is an unknown parameter of interest and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. In this example, parameter n reflects accumulation of information and $\epsilon = \epsilon_n = n^{-1/2} \to 0$ as $n \to \infty$. The loss function is taken to be $L(\hat\theta,\theta) = \|\hat\theta - \theta\|^2$, where $\|\cdot\|$ denotes the usual ℓ_2-norm, so that the quality of an estimator $\hat\theta$ is measured by the ℓ_2-norm risk $R^2(\hat\theta,\theta) = R^2_n(\hat\theta,\theta) \triangleq \mathbb{E}_\theta L(\hat\theta,\theta) = \mathbb{E}_\theta\|\hat\theta-\theta\|^2$.

Consider a family of estimators $\hat\Theta = \hat\Theta(\mathcal{I}) = \{\hat\theta(I), I \in \mathcal{I}\}$ indexed by the *cut-off* I:

$$\hat\theta(I) = (\hat\theta_i(I), i \in \mathbb{N}), \quad \hat\theta_i(I) = Y_i 1\{i \le I\}, \qquad I \in \mathcal{I} = \mathbb{N}_0.$$

We can easily check that

$$R^2(\hat\theta(I),\theta) = \mathbb{E}_\theta L(\hat\theta(I),\theta) = \mathbb{E}_\theta\|\hat\theta(I)-\theta\|^2 = \mathbb{E}_\theta\Big[\sum_{i>I}\theta_i^2 + \sum_{i\in[I]}\frac{\xi_i^2}{n}\Big] = \sum_{i>I}\theta_i^2 + \frac{I}{n}.$$

For each $\theta \in \ell_2$, there exists the best cut-off $I_o = I_o(\theta) = I_o(\theta,n)$ (if not unique, take any minimizer), called *oracle cut-off*, corresponding to the fastest risk, what we defined as *oracle rate*,

$$r^2(\theta) = r^2_n(\theta) = \inf_{I\in\mathcal{I}} R^2(\hat\theta(I),\theta) = R^2(\hat\theta(I_o),\theta) = \sum_{i>I_o}\theta_i^2 + \frac{I_o}{n}.$$

Later, in Section 5.5.1, for a penalized cut-off selector $\hat I = \hat I(Y)$ (or as a result of an empirical Bayes procedure), we establish (1.5) (hence, also (1.4)) for this example with the above oracle rate $r^2(\theta)$ and $p(\epsilon) = \frac{1}{n}$.

Introduce the Sobolev ellipsoids: for $\beta, Q > 0$,

$$\Theta_\beta = \Theta_\beta(Q) = \{\theta \in \ell_2 : \textstyle\sum_{i\in\mathbb{N}} i^{2\beta}\theta_i^2 \le Q\}.$$

It is well known that the corresponding minimax rate is $r^2(\Theta_\beta) \asymp n^{-2\beta/(2\beta+1)}$; see, for example, [12] or [71]. Further, note that $p(\epsilon) = \frac{1}{n} \ll r^2(\Theta_\beta)$ as $n \to \infty$. Assume that $\theta \in \Theta_\beta$. Let us demonstrate that in this example the local result (1.4) (or (1.5)) implies the minimax result (1.7) (or (1.8)). We only need to show (1.6). By taking $I_0 = I_0(\beta) = \lfloor n^{1/(2\beta+1)}\rfloor$, (1.6) follows:

$$\begin{aligned}\sup_{\theta\in\Theta_\beta} r^2(\theta) &= \sup_{\theta\in\Theta_\beta}\Big(\sum_{i>I_o}\theta_i^2 + \frac{I_o}{n}\Big) \le \sup_{\theta\in\Theta_\beta}\sum_{i>I_0}\frac{i^{2\beta}\theta_i^2}{I_0^{2\beta}} + \frac{I_0}{n}\\ &\le \frac{I_0}{n} + \frac{Q}{I_0^{2\beta}} \lesssim n^{-2\beta/(2\beta+1)} \asymp r^2(\Theta_\beta).\end{aligned}$$

Then the last relation and (1.4) (or (1.5)) imply (1.7) (or (1.8)).

1

1.3. BAYESIAN APPROACH

As compared to the *frequentist* paradigm, the Bayesian approach is based on the belief that there is no fixed underlying truth θ, but rather that the parameter θ is random itself and its value is realized from a prior distribution. Precisely, consider a probability measure Π on the parameter space Θ, called *prior* distribution, and a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Then the measure $\mathbb{P}_\theta$ is regarded as conditional distribution of data Y given the parameter θ. Combining the prior distribution Π on Θ and the conditional distribution $\mathbb{P}_\theta$ of the data Y given the parameter θ leads to a *posterior* distribution $\Pi(\cdot|Y)$. The core concept of the Bayesian paradigm is this posterior distribution, which can be used to make an inference on the parameter θ.

Now let us apply a Bayesian approach to the adaptation problem. There are two well known adaptive Bayesian techniques: hierarchical (full) Bayes and empirical Bayes.

According to the *hierarchical* Bayesian approach, let Π_s be a conditional distribution of θ given s and λ be a prior distribution on s. Thus, we design a two-level hierarchical prior Π on the pair (θ, s): $\theta|s \sim \Pi_s, s \sim \lambda$. The resulting hierarchical prior on θ is a mixture of the appropriate priors indexed by the structural parameter $s \in \mathcal{S}$. This and Bayes formula lead to the posterior distributions $\Pi(\theta|Y)$ and $\Pi(s|Y)$. Then all the inference on θ and s is based on these two posterior distributions. For instance, for appropriately chosen priors the full Bayes method can obtain adaptive rate-optimal recovery of the underlying truth.

Contrary to the hierarchical Bayesian approach, in the *empirical Bayes* method the unknown structural parameter s in prior Π_s is not random. The main idea of this approach is to estimate s by $\hat{s} = \hat{s}(Y)$, which is based on maximizing the marginal Bayesian likelihood function. Then we can make a statistical inference using the selected model and our standard method as in the single model situation. Thus, the empirical Bayes approach can be considered as an approximation of the hierarchical Bayes method wherein the structural parameter s is substituted by its most plausible value, instead of being integrated out. Interestingly, there is a connection of the empirical Bayes approach with penalized estimators: the marginal likelihood can be related to the penalization method (see Chapters 2, 3 and 5 for details). The empirical Bayes technique is also often computationally convenient and therefore applied not only in theory, but also in practice. In this thesis we apply the empirical Bayes method combined with a normal likelihood, whereas the true model does not have to be normal (and independence of the random errors is not required either), but only satisfying some mild *exchangeable exponential moment condition* (see Condition (A1) in Section 1.5 for details). Thus, we intend to make a *robust* inference on the unknown parameter of interest using empirical Bayes technique.

1.3.1. POSTERIOR CONTRACTION RATE

From now on, in the posteriors for θ, we will use the variable ϑ to distinguish it from the "true" θ. One of the important properties of the posterior distribution is the rate of contraction to the truth. To be more precise, a Bayesian procedure with the posterior $\Pi(\vartheta|Y)$ has a *posterior contraction rate* r_ϵ^2 if for $M > 0$ large enough

$$\Pi(L(\vartheta, \theta) \geq M r_\epsilon^2 | Y) \to 0 \quad as \quad \epsilon \to 0 \tag{1.9}$$

in $\mathbb{P}_\theta$-probability. We will also use the terms *posterior convergence rate* or *posterior concentration rate*. The contraction rate r_ϵ^2 quantifies how quickly the mass of the posterior distribution concentrates around the parameter of interest θ. Ideally, we would like (1.9) to hold for the smallest possible r_ϵ^2, preferably uniformly over $\theta \in \Theta_s$, with set $\Theta_s \subseteq \Theta$ as big as possible. Usually a global quantity (e.g., minimax rate defined by (1.1)) is taken as the posterior contraction rate $r_\epsilon^2 = r_\epsilon^2(\Theta_s)$ and (1.9) holds uniformly over $\theta \in \Theta_s$.

Now we introduce a local version of posterior contraction rate $r_\epsilon^2 = r_\epsilon^2(\theta)$, which is now allowed to depend on θ; for example, oracle rate defined by (1.3). The Bayesian procedure Π is said to have a *local posterior contraction rate* $r_\epsilon^2(\theta)$ if

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \Pi(L(\vartheta,\theta) \ge M r_\epsilon^2(\theta) | Y) \le \phi(M), \tag{1.10}$$

for some monotonically decreasing to zero function $\phi(M)$ as $M \to \infty$.

Similar to the estimation relation (1.5), the exponential non-asymptotic concentration bound for local posterior contraction rate can be defined as follows: there exist positive constants M_0, H_0, m_0 such that for any $M \ge 0$,

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \Pi\big(L(\vartheta,\theta) \ge M_0 r_\epsilon^2(\theta) + M p(\epsilon) | Y\big) \le H_0 e^{-m_0 M}, \tag{1.11}$$

where $r_\epsilon^2(\theta)$ is some local rate (e.g., of the oracle type $r^2(\theta)$ introduced by (1.3)) and $p(\epsilon)$ is some penalty term (typically $p(\epsilon) \ll r_\epsilon^2(\theta)$ as $\epsilon \to 0$). Assume the local rate $r_\epsilon^2(\theta)$ from (1.11) *covers* the scale $\{\Theta_s, s \in \mathcal{S}\}$ in the sense that

$$r_\epsilon^2(\theta) \le C r_\epsilon^2(\Theta_s) \quad \text{for all} \quad \theta \in \Theta_s, \ \ s \in \mathcal{S}, \tag{1.12}$$

where $r_\epsilon^2(\Theta_s)$ is the minimax rate over the set $\Theta_s \subseteq \Theta$, defined by (1.1). Besides, let $\frac{r_\epsilon^2(\Theta_s)}{p(\epsilon)} \to \infty$ as $\epsilon \to 0$, which is typically the case because usually $p(\epsilon) \ll r_\epsilon^2(\theta) \lesssim r_\epsilon^2(\Theta_s)$. Then the local claim (1.11) implies the global relation (1.9) with the minimax rate $r_\epsilon^2 = r_\epsilon^2(\Theta_s)$, uniformly in $\theta \in \Theta_s$. Indeed, using (1.11) and (1.12), we derive that for any $M \ge 2M_0 C$,

$$\begin{aligned}
\sup_{\theta\in\Theta_s} \mathbb{E}_\theta \Pi\big(L(\vartheta,\theta) \ge M r_\epsilon^2(\Theta_s) | Y\big) &\le \sup_{\theta\in\Theta_s} \mathbb{E}_\theta \Pi\big(L(\vartheta,\theta) \ge \tfrac{M}{2C} r_\epsilon^2(\theta) + \tfrac{M}{2} r_\epsilon^2(\Theta_s) | Y\big) \\
&\le \sup_{\theta\in\Theta_s} \mathbb{E}_\theta \Pi\big(L(\vartheta,\theta) \ge M_0 r_\epsilon^2(\theta) + \tfrac{M}{2} r_\epsilon^2(\Theta_s) | Y\big) \le H_0 e^{-m_0 M r_\epsilon^2(\Theta_s)/(2p(\epsilon))} \to 0,
\end{aligned}$$

as $\epsilon \to 0$. This means that the local result (1.11) is stronger than any global adaptive minimax result (1.9) with the minimax rate $r_\epsilon^2 = r_\epsilon^2(\Theta_s)$ over all scales which are contained in the space Θ and covered by the local rate $r_\epsilon^2(\theta)$, i.e., for which (1.12) is fulfilled.

Suppose $p(\epsilon) \le c r_\epsilon^2(\theta)$ (which is typically the case). Then this and the exponential bound (1.11) imply (1.10). Indeed, by taking $M = (M' - M_0) r_\epsilon^2(\theta)/p(\epsilon) \ge (M' - M_0) c^{-1}$, we obtain that

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathbb{E}_\theta \Pi\big(L(\vartheta,\theta) \ge M' r_\epsilon^2(\theta) | Y\big) &= \sup_{\theta\in\Theta} \mathbb{E}_\theta \Pi\big(L(\vartheta,\theta) \ge M_0 r_\epsilon^2(\theta) + M p(\epsilon) | Y\big) \\
&\le H_0 e^{-m_0 M} \le H_0 e^{-m_0 c^{-1}(M' - M_0)} \to 0,
\end{aligned}$$

as $M' \to \infty$. Moreover, (1.11) and (1.12) trivially imply the following adaptive minimax result: for any $M \geq 0$

$$\sup_{\theta \in \Theta_s} \mathbb{E}_\theta \Pi\big(L(\vartheta, \theta) \geq M_0 C r_\epsilon^2(\Theta_s) + M p(\epsilon) | Y\big) \leq H_0 e^{-m_0 M}. \tag{1.13}$$

Furthermore, the non-asymptotic exponential bounds in terms of the constant M from (1.11) and (1.13) allow to consider various asymptotic regimes depending on the concrete model (sample size $N \to \infty$, noise variance $\sigma \to 0$, or their combination) as we can let M depend in any way on N, σ, or both. Due to the above mentioned advantages of non-asymptotic exponential bounds, in this thesis we mostly focus on relations of type (1.11) when addressing the posterior contraction rate problem.

The first fundamental results on posterior contraction rates in global minimax setting were obtained in [43], [44], [81], [45] and [87]. As compared to the growing number of general results on posterior contraction rates in global setting, in the meantime a local theory on posterior contraction rates is not well developed and studied. In this thesis we are going to fill this gap. Namely, we will provide a unified approach to derive local posterior contraction rate results for many different examples of nonparametric and high-dimensional models under misspecification using empirical Bayes method. Furthermore, as we demonstrated in this section, global adaptive minimax results on posterior contraction problems over various scales will follow from our local results.

1.4. UNCERTAINTY QUANTIFICATION

In Section 1.2 we have formulated the estimation problem of unknown parameter θ in the minimax and oracle settings. However, even an optimal estimator $\hat{\theta}$ does not provide any information about its closeness to the truth. In other words, the optimal estimator $\hat{\theta}$ can be close to, or far away from the true parameter θ. It is desirable to have some sort of quantification of uncertainty of the estimator, which can be cast into the problem of constructing confidence sets for the parameter of interest.

Let us introduce the optimality framework for uncertainty quantification. Assume that Θ is a subset of a normed linear space. In this thesis, the size of a confidence set is measured by the smallest radius of a ball containing this set, hence it suffices to consider confidence balls. For the usual norm $\|\cdot\|$ in Θ, a random ball in Θ is $B(\hat{\theta}, \hat{r}) = \{\theta \in \Theta : \|\hat{\theta} - \theta\| \leq \hat{r}\}$, where the center $\hat{\theta} = \hat{\theta}(Y) : \mathcal{Y} \mapsto \Theta$ and radius $\hat{r} = \hat{r}(Y) : \mathcal{Y} \mapsto \mathbb{R}_+ = [0, +\infty]$ are measurable functions of the data Y. The goal is to construct such a confidence ball $B(\hat{\theta}, C\hat{r})$ that for any $\alpha_1, \alpha_2 \in (0, 1]$ and some functional $R(\theta) = R_{\sigma,N}(\theta)$, $R : \Theta \mapsto \mathbb{R}_+$, there exist $C, c > 0$ such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, C\hat{r})\big) \leq \alpha_1, \quad \sup_{\theta \in \Theta_1} \mathbb{P}_\theta\big(\hat{r} \geq cR(\theta)\big) \leq \alpha_2, \tag{1.14}$$

for some $\Theta_0, \Theta_1 \subseteq \Theta$. The function $R(\theta)$, called the *radial rate*, is a benchmark for the effective radius of the confidence ball $B(\hat{\theta}, C\hat{r})$. The first expression in (1.14) is called *coverage relation* and the second *size relation*. Notice that our approach is local (and hence genuinely adaptive) as the radial rate $R(\theta)$ is a function of the "true" parameter θ. The minimax adaptive version of (1.14) for a scale $\{\Theta_s, s \in \mathcal{S}\}$ (indexed by a structural parameter $s \in \mathcal{S}$, e.g., smoothness or sparsity) would be obtained by taking $\Theta_0 = \Theta_1 = \Theta_s$ and the

global radial rate $R(\theta) = r(\Theta_s)$, for all $\theta \in \Theta_s$, $s \in \mathcal{S}$, where $r^2(\Theta_s) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta_s} \mathbb{E}_\theta \|\tilde{\theta} - \theta\|^2$ is the minimax estimation rate over the sets Θ_s.

Coming back to our local framework (1.14), it is desirable to find the smallest $R(\theta)$ and the biggest Θ_0, Θ_1, for which (1.14) holds. These are contrary requirements, so we have to trade them off against each other. There are different ways of doing this, leading to different optimality frameworks. A traditional optimality framework commonly pursued in the literature (in earlier papers on the topic) was to insist on $\Theta_0 = \Theta$ in (1.14), i.e., considered confidence sets (called "honest" in some papers) must satisfy the coverage property uniformly over the entire space. Then one tries to find a "honest" confidence set with the fastest radial rate $R(\theta)$ and the biggest set Θ_1 (preferably $\Theta = \mathbb{R}^N$). However, it turned out that pursuing such an optimality framework often leads to discarding many good procedures and optimality of uninteresting ones. Many "good" confidence sets are not "honest", therefore cannot be optimal and effectively excluded from the consideration. Besides, the results of [60], [7] (formulated for the high-dimensional setting $\Theta = \mathbb{R}^N$ in the "signal+noise" model) say basically that the radial rate $R(\theta)$ cannot be of a faster order than $\sigma N^{1/4}$ for every θ and is at least of the order $\sigma N^{1/2}$ for some θ, where σ^2 is the variance of the additive noise. This means that, in the situations when the targeted optimal size $r(\theta)$ (local oracle or global minimax $r(\Theta_s)$) can be of a smaller order than $\sigma N^{1/4}$ for some θ's (which is typically the case, e.g., for smoothness and sparsity structures), this optimal size cannot be attained in the size relation uniformly over Θ and necessarily $R(\theta) \gg r(\theta)$ for some $\theta \in \Theta'$. Thus, insisting on $\Theta_0 = \Theta$ implies that either the radial rate $R(\theta)$ or the set Θ_1 in the size relation has to be sacrificed: $\Theta_1 = \Theta$ but $R(\theta) \gg r(\theta)$ for $\theta \in \Theta'$, or $R(\theta) = r(\theta)$ but $\Theta_1 = \Theta \backslash \Theta'$.

Another, seemingly more reasonable approach to optimality developed recently in the literature is to sacrifice in the set $\Theta_0 = \Theta \backslash \Theta_{\text{dec}}$ by removing a preferably small portion of "deceptive parameters" Θ_{dec} from Θ in the coverage property, so that the size property would then hold with $R(\theta) = r(\theta)$ uniformly over $\Theta_1 = \Theta$. To summarize, there are situations when the overall uniform coverage and optimal size properties cannot hold together and it is necessary to sacrifice at least one of these. This is the core of the so called *deceptiveness* issue in the uncertainty quantification problem. This deceptiveness phenomenon is well understood only for smoothness and sparsity structures; see [9], [13] and further references therein. Typically, global radial rates for specific smoothness (or sparsity) structures are studied in the literature. Local results, delivering also (global) adaptive minimax results for smoothness and sparsity structures, are obtained in [9] and [13].

1.5. GENERAL FRAMEWORK OF PROJECTION STRUCTURES

In the previous section we introduced the uncertainty quantification problem. To the best of our knowledge, there is no general robust framework on uncertainty quantification in a local setting. In this thesis, namely in Chapter 5, we attempt to address this issue. In this section, we introduce a general framework of so called *projection structures* and equip it with necessary notions and conditions which will be used later in Chapter 5. The largest part of this section is contained in paper [17].

1

Suppose we observe a random element $(Y, X) \in (\mathcal{Y} \times \mathcal{X})$:

$$Y \sim \mathbb{P}_\theta = \mathbb{P}_{\theta,X},\ \theta \in \Theta \subseteq \mathcal{Y}, \quad \text{such that} \quad \mathbb{E}_\theta Y = \theta(X) = \theta,$$

where $\mathbb{P}_\theta$ is the probability measure of Y ($\mathbb{E}_\theta$ is the corresponding expectation) depending on an unknown high-dimensional parameter of interest θ. By default, $\Theta = \mathcal{Y} = \mathbb{R}^N$, $\mathcal{X} = \mathbb{R}^{d_X}$ for "big" $N, d_X \in \mathbb{N}$ (with the usual norm $\|\cdot\|$), unless stated otherwise. In some particular models (see Section 5.5), $\mathcal{Y}, \Theta \subseteq \mathbb{R}^\infty$ can be infinite dimensional and Θ can be a proper subset of $\mathcal{Y}$, e.g., $\mathcal{Y} = \mathbb{R}^\infty$ and $\Theta = \ell_2$. Those particular models can also be reduced to the high-dimensional case by assuming that any $\theta \in \Theta \subseteq \mathbb{R}^\infty$ can be arbitrarily well approximated by $\tilde{\theta} \in \mathbb{R}^N$ for sufficiently large $N \in \mathbb{N}$.

Useful inference in high-dimensional models is not possible without some (approximate) structure on the parameter of interest, the basic idea is to reduce the "effective" dimensionality of the high-dimensional θ. The most popular structural assumptions are *smoothness*, *sparsity* and *clustering*. These structures and many others can be represented via appropriate families of linear spaces $\mathbb{L}_I \subseteq \mathcal{Y}$, $I \in \mathcal{I}$. Precisely, we introduce a finite (or countable) family $\mathcal{I}$ of possible structures I and an associated family of linear subspaces $\{\mathbb{L}_I, I \in \mathcal{I}\}$ of $\mathcal{Y}$, which express these structures. This in turn determines the family of corresponding projection operators $\{\mathrm{P}_I, I \in \mathcal{I}\}$ onto linear subspaces $\{\mathbb{L}_I, I \in \mathcal{I}\}$. The true θ is "approximately structured" according to the family $\mathcal{I}$ if $\|\theta - \mathrm{P}_{I^*}\theta\|^2 = \min_{I \in \mathcal{I}} \|\theta - \mathrm{P}_I\theta\|^2$ is close to zero. If $\|\theta - \mathrm{P}_{I^*}\theta\|^2 = 0$, the true θ happens to be exactly structured, i.e., $\theta \in \mathbb{L}_{I^*}$ and I^* has the meaning of the "true structure" of the true θ.

Let $\sigma\xi = Y - \mathbb{E}_\theta Y$ (any Y is its expectation plus zero mean "noise"), $\sigma > 0$ be the "noise intensity", then

$$Y = \theta(X) + \sigma\xi = \theta + \sigma\xi \quad \text{with} \quad \mathbb{E}_\theta \xi = 0. \tag{1.15}$$

The (known) parameter σ is introduced to accommodate certain asymptotic regimes in particular models, $\sigma \to 0$ reflects an information increase. How the underlying measure $\mathbb{P}_{\theta,X}$ of Y (and $\theta = \theta(X) = \mathbb{E}_\theta Y$) depends on X will only be exploited when constructing specific families of *structures* $\mathcal{I} = \mathcal{I}(X)$ (e.g., in the linear regression model from Section 5.5.7). We skip the dependence of structures on X in further notation.

The goal is to make inference on the parameter θ based on the data Y: recovery of θ (*estimation* and *posterior contraction*), *structure recovery*, and *uncertainty quantification* by constructing an *optimal confidence set*. We pursue *local* inference in the sense that no structure on θ is a priori assumed but we aim to extract as much structure (according to the family of structures $\mathcal{I}$, once $\mathcal{I}$ is chosen) as there is in the underlying θ. We will make this notion precise later. We also pursue *robust* inference in the sense that the distribution of ξ is unknown and can depend on θ (often we suppress this dependence in notation), the coordinates ξ_i's of ξ do not have to be iid, even not independent. The distribution of ξ is assumed to satisfy only certain mild condition; see Condition (A1) in Section 1.5.2. We derive *non-asymptotic* results, which imply asymptotic assertions if needed. Possible asymptotic regimes are: high-dimensional setup $N \to \infty$ (the leading case in the literature for high-dimensional models $\mathcal{Y} = \mathbb{R}^N$), decreasing noise level $\sigma \to 0$, or their combination, e.g., $\sigma = N^{-1/2}$ and $N \to \infty$.

1.5.1. SLICING INTO STRUCTURAL LAYERS, LAYER COMPLEXITY

As we mentioned before, useful inference is not possible without some structure on parameter θ. The structure on each $\theta \in \Theta$ is represented by the slicing $\Theta \subseteq \cup_{I\in\mathcal{I}} \mathbb{L}_I$. Indeed, then for any $\theta \in \Theta$ there exists an $I \in \mathcal{I}$ such that $\theta \in \mathbb{L}_I$, this I (and $\mathbb{L}_I$) has the meaning of the structure of that θ. Structure I of θ is always determined via the corresponding linear space $\mathbb{L}_I \ni \theta$, so by saying "I is the structure of θ" we mean that $\theta \in \mathbb{L}_I$ (or can be well approximated by $\mathrm{P}_I\theta$).

Remark 1.1. *There may be $\mathbb{L}_I = \mathbb{L}_{I'}$ for different $I, I' \in \mathcal{I}$, in other words, the family of structures $\mathcal{I}$ can have redundancy. Without loss of generality, we could assume that the family $\mathcal{I}$ is "cleaned up" in the sense that each subspace $\mathbb{L} \in \mathcal{L}_\mathcal{I}$ is represented in $\mathcal{I}$ by only one (arbitrary) element $J = J(\mathbb{L})$ from the set $\{I \in \mathcal{I} : \mathbb{L}_I = \mathbb{L}\}$. Mathematically, this means that the resulting "cleaned up" family $\mathcal{I}$ of structures consists of equivalence classes on the original collection of all structures with the equivalence relation: $I_1 \sim I_2$ if and only if $\mathbb{L}_{I_1} = \mathbb{L}_{I_2}$, so that $|\mathcal{I}| = |\mathcal{L}_\mathcal{I}|$ in this case.*

However, in general $|\mathcal{L}_\mathcal{I}| \le |\mathcal{I}|$ and this redundancy can be beneficial in some practical situations when searching (or optimizing in an inference procedure) over a possibly redundant family of structures $\mathcal{I}$ can be described and realized easier than over the "cleaned up" version of it. The only price for this redundancy is a bigger sum (because of more terms) in Condition (A2), resulting in a bigger constant C_v in Condition (A2) for a redundant $\mathcal{I}$. In many situations this is a mild price, see for example Section 5.5.7.

Clearly, θ can have many structures, and we would like to distinguish the "simplest" structure of θ. For that, we introduce a measure of structure complexity below. Suppose the following condition is fulfilled for the vector ξ from (1.15): for some $\alpha > 0$ and $(d_I)_{I\in\mathcal{I}}$ with $d_I \ge 0$,

$$\mathbb{E}_\theta \exp\left\{\alpha \|\mathrm{P}_I \xi\|^2\right\} \le \exp(d_I) \quad \text{for all } I \in \mathcal{I},\ \theta \in \Theta. \tag{A0}$$

Remark 1.2. *It is desirable to have the bound (A0) in the tightest possible form, by determining the smallest sequence $(d_I)_{I\in\mathcal{I}}$ for which (A0) holds with a given $\alpha > 0$. (A0) is useful only when all the d_I's are finite. (Notice that (A0) always holds for any $\alpha > 0$, if the d_I's are allowed to be infinite.) Thus, instead of (A0), we could equivalently assume $\sup_{\theta\in\Theta} \mathbb{E}_\theta \exp\left\{\alpha \|\mathrm{P}_I \xi\|^2\right\} < \infty$ for all $I \in \mathcal{I}$. Then the smallest d_I's for which (A0) holds are*

$$d_I = \sup_{\theta\in\Theta} \log\left(\mathbb{E}_\theta \exp\left\{\alpha \|\mathrm{P}_I \xi\|^2\right\}\right), \quad I \in \mathcal{I}. \tag{1.16}$$

The quantity d_I can be seen as statistical dimension *of structure I, reflecting in a way the* complexity *of the structure I (space $\mathbb{L}_I$): the bigger d_I, the more complex the structure I. Notice that if the distribution of ξ does not depend on θ, then there is no $\sup_{\theta\in\Theta}$ in (1.16). Typically, in such cases $d_I \asymp \dim(\mathbb{L}_I)$. The bound (A0) holds, for example, for standard normal ξ with $\alpha = 0.43$ and $d_I = \dim(\mathbb{L}_I)$; see Remark 1.6 below.*

A subfamily $\mathcal{J} \subseteq \mathcal{I}$ of structures is called *structural layer* (or just layer) in $\mathcal{I}$. By $\mathcal{L}_\mathcal{J} = \{\mathbb{L}_I, I \in \mathcal{J}\}$ we denote the corresponding layer in the family of all linear subspaces $\mathcal{L}_\mathcal{I} = \{\mathbb{L}_I, I \in \mathcal{I}\}$. Now we characterize the *complexity of a layer* $\mathcal{J} \subseteq \mathcal{I}$. A version of this

1

key notion (in a different context) is present in [42]. Let us give a simple (but important) heuristics for controlling the maximal projected error $\max_{I\in\mathcal{J}}\|\mathrm{P}_I\xi\|^2$ over a layer $\mathcal{J}$. Denote $d_{\mathcal{J}}=\max_{I\in\mathcal{J}} d_I$. Using (A0) and Jensen's inequality, we derive

$$\exp\{\alpha\mathbb{E}_\theta \max_{I\in\mathcal{J}}\|\mathrm{P}_I\xi\|^2\}=\exp\{\alpha\mathbb{E}_\theta \max_{\mathbb{L}_I\in\mathcal{L}_{\mathcal{J}}}\|\mathrm{P}_I\xi\|^2\}\le\mathbb{E}_\theta\exp\{\alpha \max_{\mathbb{L}_I\in\mathcal{L}_{\mathcal{J}}}\|\mathrm{P}_I\xi\|^2\}$$
$$\le\sum_{\mathbb{L}_I\in\mathcal{L}_{\mathcal{J}}}\mathbb{E}_\theta e^{\alpha\|\mathrm{P}_I\xi\|^2}\le e^{d_{\mathcal{J}}+\log|\mathcal{L}_{\mathcal{J}}|}. \tag{1.17}$$

Hence, under (A0) we control the maximal projected error $\max_{I\in\mathcal{J}}\|\mathrm{P}_I\xi\|^2$ up to the order of $d_{\mathcal{J}}+\log|\mathcal{L}_{\mathcal{J}}|$. It is this sum that characterizes the complexity of the layer $\mathcal{J}$ (layer $\mathcal{L}_{\mathcal{J}}$). Define the complexity of the layer $\mathcal{J}$ (layer $\mathcal{L}_{\mathcal{J}}$) as

$$c(\mathcal{J})\triangleq d_{\mathcal{J}}+\log|\mathcal{L}_{\mathcal{J}}|\le d_{\mathcal{J}}+\log|\mathcal{J}|,\quad\text{where } d_{\mathcal{J}}=\max_{I\in\mathcal{J}}d_I. \tag{1.18}$$

The second equality holds because in general $|\mathcal{L}_{\mathcal{J}}|\le|\mathcal{J}|$ in view of Remark 1.1.

Next, introduce a surjective function $s:\mathcal{I}\mapsto\mathcal{S}$, for some set $\mathcal{S}$, called the *structural slicing mapping*. This function slices the family $\mathcal{I}$ in layers

$$\mathcal{I}_s=\{I\in\mathcal{I}: s(I)=s\},\quad s\in\mathcal{S},$$

i.e., $\mathcal{I}=\cup_{s\in\mathcal{S}}I_s$, $\mathcal{S}$ marks the collection of all layers $\mathcal{I}_s$. Clearly, the structure I belongs to the layer $\mathcal{I}_{s(I)}$, and any slicing of $\mathcal{I}$ can be realized by appropriate function $s(I)$, $I\in\mathcal{I}$. This also leads to the corresponding slicing of the parameter space $\Theta\subseteq\cup_{s\in\mathcal{S}}\mathbb{L}_{\mathcal{I}_s}$. The quantity $s(I)$ typically describes some features of the space $\mathbb{L}_I$. For example, $s(I)$ can be the dimension (or some function of it) of $\mathbb{L}_I$.

According to (1.18) and Remark 1.1, the complexity of the layer $\mathcal{I}_s$ is then

$$c(s)=d_s+\log|\mathcal{L}_s|\le d_s+\log|\mathcal{I}_s|,\quad s\in\mathcal{S}, \tag{1.19}$$

where, slightly abusing notation, we adopted the following notional conventions for brevity: $c(s)=c(\mathcal{I}_s)$, $d_s=d_{\mathcal{I}_s}=\max_{I\in\mathcal{I}_s}d_I$, $\mathcal{L}_s=\mathcal{L}_{\mathcal{I}_s}=\{\mathbb{L}_I, I\in\mathcal{I}_s\}$, with $c(\mathcal{J})$ defined by (1.18).

1.5.2. CONDITIONS

Recall that we have introduced a structural slicing mapping $s:\mathcal{I}\mapsto\mathcal{S}$, which slices the family of structures $\mathcal{I}$ and the parameter space Θ in structural layers: $\mathcal{I}=\cup_{s\in\mathcal{S}}I_s$ and $\Theta\subseteq\cup_{s\in\mathcal{S}}\mathbb{L}_s$, $s\in\mathcal{S}$, where $\mathcal{I}_s=\{I\in\mathcal{I}: s(I)=s\}$ and $\mathbb{L}_s=\{\mathbb{L}_I: s(I)=s\}$.

In previous section we proposed condition (A0) which led to the notion (1.18) of layer complexity $c(\mathcal{J})$. We can relax condition (A0) by imposing it on the layers $\mathcal{I}_s$ only (instead of all $I\in\mathcal{I}$). Precisely, we impose the following so called *exchangeable exponential moment condition* on the random vector ξ from (1.15).

CONDITION (A1). For some structural slicing mapping $s:\mathcal{I}\mapsto\mathcal{S}$, sequence $(d_s)_{s\in\mathcal{S}}$ and $\alpha>0$,

$$\mathbb{E}_\theta\exp\{\alpha\|\mathrm{P}_I\xi\|^2\}\le\exp\{d_{s(I)}\},\quad I\in\mathcal{I},\ \theta\in\Theta. \tag{A1}$$

Without loss of generality, assume $\alpha\in(0,1]$ and $d_{s(I)}\gtrsim\dim(\mathbb{L}_I)$, $I\in\mathcal{I}$.

Remark 1.3. *Note that there is no need to assume* $d_{s(I)} \gtrsim \dim(\mathbb{L}_I)$, $I \in \mathcal{I}$, *as all the results below hold true (with small adjustments) for any* $(d_s)_{s\in\mathcal{S}}$ *for which* (A1) *is fulfilled. The only place where this is used is the relation* (5.32)*, and this can easily be fixed by modifying the empirical Bayes posterior* (5.10) *for* I *(use* $d_{s(I)}$ *instead of* $\dim(\mathbb{L}_I)$ *in* (5.10)*).*

Remark 1.4. *Note that if* $\mathbb{E}_\theta \exp\{\beta\|\mathrm{P}_I\xi\|^2\} \le \exp\{Bd_{s(I)}\}$ *is fulfilled for some* $\beta > 0$ *and* $B > 1$*, then it implies Condition* (A1)*. Indeed, by taking* $\alpha = \frac{\beta}{B}$ *and using the fact that* $\mathbb{E}|Z| \le (\mathbb{E}|Z|^{\frac{\beta}{\alpha}})^{\frac{\alpha}{\beta}}$ *for any random variable* Z*, we have that*

$$\mathbb{E}_\theta e^{\alpha\|\mathrm{P}_I\xi\|^2} = \mathbb{E}_\theta e^{(\beta\|\mathrm{P}_I\xi\|^2)\frac{\alpha}{\beta}} \le \left(\mathbb{E}_\theta e^{\left((\beta\|\mathrm{P}_I\xi\|^2)\frac{\alpha}{\beta}\right)\frac{\beta}{\alpha}}\right)^{\frac{\alpha}{\beta}} = \left(\mathbb{E}_\theta e^{\beta\|\mathrm{P}_I\xi\|^2}\right)^{\frac{\alpha}{\beta}} \le e^{B\alpha\beta^{-1}d_{s(I)}} = e^{d_{s(I)}}.$$

Remark 1.5. *In view of Remark* 1.2*,* (A1) *is equivalent to the following condition: for* $\alpha > 0$*,*

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \exp\{\alpha\|\mathrm{P}_I\xi\|^2\} < \infty, \quad I \in \mathcal{I}. \tag{A1'}$$

Then the smallest d_s*'s for which* (A1) *is fulfilled are* $d_s = \max_{I\in\mathcal{I}_s} d_I$*, where* d_I *is defined by* (1.16)*. Referring to Remark* 1.2*,* d_s *can be interpreted as* statistical dimension *of the layer* $\mathcal{I}_s$*, which is the first term of its total complexity* (1.19) *.*

Remark 1.6. *Condition* (A1) *holds for high-dimensional independent normal* ξ_i*'s with* $d_{s(I)} = \dim(\mathbb{L}_I)$*, irrespective of the linear spaces* $\mathbb{L}_I$*,* $I \in \mathcal{I}$*. In fact, the stronger Condition* (A0) *holds with* $d_I = \dim(\mathbb{L}_I)$*. Indeed, if* $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$*,* $\|\mathrm{P}_I\xi\|^2 \sim \chi^2_{\dim(\mathbb{L}_I)}$*, the chi-squared distribution with* $d_I = \dim(\mathbb{L}_I)$ *degrees of freedom. Hence, for any* $t < \frac{1}{2}$ *we have that* $\mathbb{E}\exp\{t\|\mathrm{P}_I\xi\|^2\} = (1-2t)^{-d_I/2}$*. Since* $(1-2t)^{-d_I/2} \le e^{d_I}$ *for any* $t \le (1-e^{-2})/2 \approx 0.432$*. By taking* $t = 0.4$*, we derive* $\mathbb{E}e^{0.4\|\mathrm{P}_I\xi\|^2} \le e^{d_I}$*. Hence, Condition* (A0) *is fulfilled with* $\alpha = 0.4$ *and* $d_I = \dim(\mathbb{L}_I)$*.*

Importantly, Condition (A1) allows quite some flexibility, which is crucial when treating concrete models and significantly broadens the range of models falling into our general framework. First, the distribution of ξ may depend on θ (note however that in many important models from Section 5.5, the distribution of ξ does not depend on θ). Besides, the coordinates ξ_i's of ξ do not have to be iid and may even be non-independent. For example, for the "signal+noise" model with the sparsity structure, it was shown in [13] that Condition (A1) is fulfilled for the ξ_i's generated according an autoregressive model.

In case of "signal+noise" model with the sparsity structure, one can relate (cf. [13]) Condition (A1) to the so called *sub-gaussianity* condition on $\xi = (\xi_i, i \in [N])$. The vector ξ is called *sub-gaussian* with parameter $\rho > 0$ if

$$\mathbb{P}(|\langle v, \xi\rangle| > t) \le e^{-\rho t^2} \text{ for all } t \ge 0,\ v \in \mathbb{R}^N \text{ such that } \|v\| = 1. \tag{1.20}$$

In this case, in [13] we showed that the sub-gaussianity condition is equivalent to Condition (A1) for independent ξ_i's; for dependent ξ_i's, the sub-gaussianity condition (1.20) and Condition (A1) are close, but in general incomparable. For example, if $\xi_i = \xi_0$, $i \in [n]$, for some bounded random variable ξ_0 (say, uniform on $[-1,1]$), then a version of Condition (A1) (for the sparsity structure) trivially holds whereas the sub-gaussianity condition is not fulfilled.

It is desirable to use a slicing $s : \mathcal{I} \mapsto \mathcal{S}$ that is parsimonious in the sense that the maximum $d_s = \max_{I \in \mathcal{I}_s} d_I$ degenerates, i.e., $d_I = d_J$ for all $I, J \in \mathcal{I}_s$, so $d_{s(I)} = d_I$. In other words, $d_I = h(s(I))$, $I \in \mathcal{I}$, for some function $h : \mathcal{S} \mapsto \mathbb{N}_0$. Since typically $d_I \asymp \dim(\mathbb{L}_I)$, in many examples of Section 5.5 we will take a slicing s such that $\dim(\mathbb{L}_I) = h(s(I))$, $I \in \mathcal{I}$, for some function h.

The complexity $c(s)$ of the layer $\mathcal{I}_s$ is defined in the previous section by (1.19). Instead, from now on we will work with a majorant of the layer complexity, some function $\rho : \mathcal{S} \mapsto \mathbb{R}_+$ such that

$$\rho(s) \geq d_s + \log|\mathcal{I}_s| \geq c(s), \quad s \in \mathcal{S}. \tag{1.21}$$

Remark 1.7. *We can use an up-to-a-constant majorant $\rho(s) \gtrsim c(s)$ instead of (1.21) by adjusting $\alpha \in (0,1]$ in (A1), but without loss of generality we stick to (1.21) for the sake of a clean mathematical exposition.*

The reasoning in (1.17) gives a heuristic explanation why later on at least a multiple of the complexity $c(s)$ must be present in the prior on I (and as penalty term in the penalization method). This quantity will also enter the local (oracle) rate. It is therefore desirable to use the smallest possible majorant $\rho(s)$ in (1.21) in order to derive stronger results. In this light, the best majorant in (1.21) is the complexity itself $\rho(s) = d_s + \log|\mathcal{I}_s| = c(s)$ for the "cleaned up" family of structures $\mathcal{I}$ (see Remark 1.1). On the other hand, in the end any majorant $\rho(s) \gtrsim c(s)$ will do the job. The reason to allow an arbitrary majorant $\rho(s)$ is that d_s and $|\mathcal{I}_s|$ (or $|\mathcal{L}_s|$) may be difficult to compute, whereas some closed form upper bounds can be derived. Of course, this comes at the price of a bigger resulting local rate because this majorant will then enter the local rate.

In the proof of Theorem 5.1 below, we will need a bound for $\left[\mathbb{E}_\theta \|\mathrm{P}_I \xi\|^4\right]^{1/2}$, for each $I \in \mathcal{I}$. Condition (A1) ensures such a bound. Indeed, since $x^2 \leq e^{2x}$ for all $x \geq 0$, by the Hölder inequality and (A1), we obtain for any $t \in (0, 1/2]$ and $I \in \mathcal{I}$,

$$\mathbb{E}_\theta \|\mathrm{P}_I \xi\|^4 \leq \frac{\mathbb{E}_\theta e^{2t\alpha\|\mathrm{P}_I\xi\|^2}}{(t\alpha)^2} \leq \frac{\left(\mathbb{E}_\theta e^{\alpha\|\mathrm{P}_I\xi\|^2}\right)^{2t}}{(t\alpha)^2} \leq \frac{e^{2t\rho(s(I))}}{(t\alpha)^2}. \tag{1.22}$$

In case $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, Condition (A0) is fulfilled with $d_I = \dim(\mathbb{L}_I)$, so that (A1) holds also with $d_s = \max_{I \in \mathcal{I}_s} d_I$. As $\|\mathrm{P}_I \xi\|^2 \sim \chi^2_{d_I}$, instead of (1.22), a better bound can be used in this case: $\left[\mathbb{E}\|\mathrm{P}_I\xi\|^4\right]^{1/2} = \left(d_I^2 + 2d_I\right)^{1/2} \leq d_I + 1 \leq \rho(s(I)) + 1$.

We finish this section by introducing the conditions on the family of structures $\mathcal{I}$ and the majorant $\rho(s)$ from (1.21), which we will need in the theorems.

CONDITION (A2). For some $\nu, C_\nu > 0$, $\sum_{I \in \mathcal{I}} e^{-\nu\rho(s(I))} \leq C_\nu$. (A2)

Remark 1.8. *Notice that in view of (1.21), for $\nu \geq 1$ we obtain the bound*

$$\sum_{I \in \mathcal{I}} e^{-\nu\rho(s(I))} = \sum_{s \in \mathcal{S}} \sum_{I \in \mathcal{I}_s} e^{-\nu\rho(s(I))} \leq \sum_{s \in \mathcal{S}} e^{-\nu d_s - (\nu-1)\log|\mathcal{I}_s|} \leq \sum_{s \in \mathcal{S}} e^{-\nu d_s}.$$

So, (A2) is satisfied if $\sum_{s \in \mathcal{S}} e^{-\nu d_s} \leq C_\nu$ for $\nu \geq 1$. Informally, this condition means that the majorant ρ is large enough to match the massiveness (reflected by the cardinalities $|\mathcal{I}_s|$, $s \in \mathcal{S}$) and the complexity (reflected by the sequence $(d_s)_{s \in \mathcal{S}}$) of the family of structures $\mathcal{I}$.

CONDITION (A3). For any $I_0, I_1 \in \mathcal{I}$ there exists $I' = I'(I_0, I_1) \in \mathcal{I}$ such that $(\mathbb{L}_{I_0} \cup \mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'}$ and $\rho(s(I')) \le \rho(s(I_0)) + \rho(s(I_1))$.

Remark 1.9. *Typically, Condition (A3) is fulfilled with $I' = I'(I_0, I_1) \in \mathcal{I}$ such that $\mathbb{L}_{I'} = \mathbb{L}_{I_0} + \mathbb{L}_{I_1}$. This is the case for almost all examples in Section 5.5.*

Let us formulate a slightly stronger version of Condition (A3) called Condition (A3'): for any $I_0, I_1 \in \mathcal{I}$ there exist $I' = I'(I_0, I_1) \in \mathcal{I}$ and $I'' = I''(I_0, I_1) \in \mathcal{I}$ such that $\mathbb{L}_{I'} = \mathbb{L}_{I_0} + \mathbb{L}_{I_1}$, $\mathbb{L}_{I''} = \mathbb{L}_{I_0} \cap \mathbb{L}_{I_1}$, $\rho(s(I')) \le \rho(s(I_0)) + \rho(s(I_1))$ and $\rho(s(I'')) \le \rho(s(I_0))$.

The constants $\alpha \in (0, 1]$ and $\nu > 0$ will be fixed throughout and we omit the dependence on these constants in all further notation.

1.5.3. EXAMPLES OF MODELS AND STRUCTURES

There are numerous examples of models and structures falling into our general framework. In this thesis we treat the following models and structures: 1) signal+noise model with smoothness structure (Sobolev ellipsoids and hyperrectangles, analytic and tail classes); 2) signal+noise model under wavelet basis (Besov balls); 3) signal+noise model with (multi-level) sparsity structure (multi-level sparsity is considered for the first time); 4) noisy function on a large graph (Laplacian graph) with smoothness structure; 5) density estimation with smoothness structure; 6) biclustering model (also for stochastic block model and graphon classes); 7) linear regression with sparsity structure, with group sparsity, with group clustering, and with mixture structure; 8) aggregation in nonparametric regression; 9) isotonic, unimodal and convex regressions; 10) dictionary learning; 11) mean matrix with submatrix sparsity; 12) covariance matrix with banding and sparsity structures. We emphasize that the scope of our approach extends further than just the above mentioned cases.

To demonstrate how the above introduced general framework of projection structures applies to concrete models and structures, here we briefly outline only three particular examples: signal+noise model with smoothness structure, signal+noise model with sparsity structure and stochastic block model. For all three examples we specify the family of structures $\mathcal{I}$, the structural slicing mapping $s : \mathcal{I} \mapsto \mathcal{S}$ and majorant $\rho(s(I))$ for the layer complexity. A more elaborate treatment of these examples and all the others listed above is provided in Section 5.5. In Section 5.5 we also verify Conditions (A1)-(A4) whenever appropriate and formulate the results on estimation, posterior contraction, weak structure recovery and uncertainty quantification.

Signal+noise model with smoothness structure. Assume that the data $Y = (Y_i)_{i \in \mathbb{N}}$ come from the model

$$Y_i = \theta_i + \tfrac{1}{\sqrt{n}} \xi_i, \quad i \in \mathbb{N},$$

where $\theta = (\theta_i)_{i \in \mathbb{N}} \in \Theta = \ell_2$ is an unknown parameter and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0, 1)$.

Admittedly, this is an infinite dimensional model as compared with the default high-dimensional setting, but also in this case all the results (given in Section 5.5.1) go through exactly in the same way with only one minor peculiarity: all the sums over $I \in \mathcal{I}$ become countably infinite instead of finite. Alternatively (although there is no need for this), one

could consider a finite dimensional model approximating the original infinite dimensional model with arbitrary accuracy.

In this case, the smoothness structure is modeled by the linear spaces

$$\mathbb{L}_I = \{x \in \ell_2 : x_i = 0 \text{ for all } i \geq I+1\}, \quad I \in \mathcal{I} = \mathbb{N}_0.$$

We have $d_I = \dim(\mathbb{L}_I) = I$, the structural slicing mapping is taken to be $s(I) = I$, so that $\mathcal{S} = \mathcal{I} = \mathbb{N}_0$ and $\mathcal{I}_{s(I)} = \{I\}$. Hence $\log|\mathcal{I}_s| = 0$ for all $s \in \mathcal{S}$, and we thus take the majorant $\rho(s(I)) = d_{s(I)} + \log|\mathcal{I}_{s(I)}| = d_I = I$.

Signal+noise model with sparsity structure. Assume that the data $Y = (Y_i)_{i\in[n]}$ come from the model

$$Y_i = \theta_i + \sigma\xi_i, \ \ i \in [n],$$

where $\theta = (\theta_i)_{i\in[n]} \in \Theta = \mathbb{R}^n$ is an unknown parameter and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. The high-dimensional vector θ is assumed to be *sparse*.

The classical sparsity structure is modeled by the linear spaces

$$\mathbb{L}_I = \{x \in \mathbb{R}^n : x_i = 0, i \in I^c\}, \quad I \in \mathcal{I} = \{J : J \subseteq [n]\}.$$

In this case, $d_I = \dim(\mathbb{L}_I) = |I|$, $\|\theta - \mathrm{P}_I\theta\|^2 = \sum_{i\in I^c}\theta_i^2$, the structural slicing mapping is defined to be $s(I) = |I| \in \mathcal{S} \triangleq [n]_0$. Compute $|\mathcal{I}_{s(I)}| = \binom{n}{|I|}$, hence $\log|\mathcal{I}_{s(I)}| = \log\binom{n}{|I|} \leq |I|\log(\frac{en}{|I|})$. Since $d_{s(I)} + \log|\mathcal{I}_{s(I)}| = d_I + \log|\mathcal{I}_{s(I)}| \leq |I| + |I|\log(\frac{en}{|I|})$, we take the majorant $\rho(s(I)) = 2|I|\log(\frac{en}{|I|})$.

Stochastic block model. Suppose we observe a matrix $Y = (Y_{ij}) \in \mathbb{R}^{n\times n}$, where

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad i, j \in [n],$$

$\theta = (\theta_{ij}) \in \mathbb{R}^{n\times n}$ is an unknown high-dimensional parameter of interest with *clustering* structure, $\sigma > 0$ is the known noise intensity. Precisely, in this case clustering structure means that the row clusters coincide with the column clusters of the matrix $\theta = (\theta_{ij}) \in \mathbb{R}^{n\times n}$, and the values θ_{ij} are the same for i, j from the same clusters. The observations Y_{ij} can be associated with network data. In this case Y_{ij} stands for the presence or absence of an edge between vertices i and j in the network interpretation.

For $k \in [n]$, consider a mapping $z : [n] \mapsto [k]$. Each mapping $z \in [k]^{[n]}$ determines (uniquely) the pertinent partition $I = I(z)$ of the rows and columns of any matrix $(M_{ij}) \in \mathbb{R}^{n\times n}$ into k^2 blocks:

$$[n] \times [n] = z^{-1}([k]) \times z^{-1}([k]) = \cup_{(I_i^1, I_j^2)\in I}(I_i^1, I_j^2),$$

where $I_i^1 = z^{-1}(i)$ and $I_j^2 = z^{-1}(j)$. So, the collection of all mappings $\mathcal{Z} = \mathcal{Z}(n) = \{z \in [k]^{[n]}, k \in [n]\}$ yields the collection of all clustering structures (which are all *clustered* partitions of $[n]$):

$$\mathcal{I} = \mathcal{I}(n) = \{I(z), z \in [k]^{[n]}, k \in [n]\}.$$

A clustering structure $I \in \mathcal{I}$ in terms of parameter θ is expressed by imposing $\theta \in \mathbb{L}_I \subseteq \mathbb{R}^{n_1 \times n_2}$, where the linear subspace $\mathbb{L}_I$ is defined as

$$\mathbb{L}_I = \left\{ x \in \mathbb{R}^{n_1 \times n_2} : x_{ij} = x_{i'j'} \ \forall\, (i,j),(i',j') \in (I_1, I_2),\ \forall\, (I_1, I_2) \in I \right\}.$$

The structural slicing mapping $s : \mathcal{I} \mapsto \mathcal{S}$ is defined as $s(I) \in [n] \triangleq \mathcal{S}$, where $s(I)$ denotes the number of nonempty row or column blocks in the structure $I \in \mathcal{I}$. Then $d_{s(I)} = d_I = \dim(\mathbb{L}_I) = s^2(I)$.

Let us propose a reasonable majorant $\rho(s)$ for the layer complexity $d_s + \log|\mathcal{I}_s| = s^2 + \log|\mathcal{I}_s|$. Clearly, $|\mathcal{I}_s| \le s^n$. So we take $\rho(s(I)) = s^2(I) + n \log s(I)$.

1.6. SCOPE OF THE THESIS

In this thesis we focus on three important statistical problems: estimation, posterior contraction rate and uncertainty quantification, by using empirical Bayes and penalization methods. We note that on the way we also obtained some interesting results on structure recovery (in a weak sense) for various models and structures. The main contribution of this thesis is development of general robust framework and approach for addressing the above mentioned inference problems and applying these to a number of various examples of high-dimensional and nonparametric models and structures under possible misspecification. Admittedly, some parts of Chapters 2, 3 and 4 are contained in Chapter 5, which should be seen as the main chapter of this thesis. The problems considered in Chapters 2 and 4 can in principle be treated via general framework (developed in Chapter 5) by applying the general framework machinery to the corresponding particular projection structures and by performing the necessary computations. However, we started the present project with the three particular models and structures from Chapters 2, 3 and 4, which are interesting and important on their own right. The main goal was to obtain novel results for the grand problem of uncertainty quantification (on the way solving other inference problems as well) in these three important models. As usually in the course of research, one moves from simple situation to more complex and more abstract settings in steps, we too studied first the three models presented by Chapters 2, 3 and 4, before we were able to generalize our approach and develop the general framework of Chapter 5. Results of Chapters 2 and 4 formed the necessary basis for development of the general framework and represent the research path we followed within this project, culminating in the general framework results of Chapter 5.

Chapter 2 is based on paper [13]. In this chapter we construct an empirical Bayes posterior in the general *signal+noise* (allowing non-normal, non-independent observations) model. Then we use this empirical Bayes posterior for *uncertainty quantification* for the unknown, possibly sparse, signal. We introduce a novel *excessive bias restriction* (EBR) condition, which gives rise to a new slicing of the entire space that is suitable for uncertainty quantification. Under EBR and some mild *exchangeable exponential moment condition* on the noise, we establish the local (oracle) optimality of the proposed confidence ball. Without EBR, we derive the full coverage for confidence balls of at least $\sigma n^{1/4}$-radius (n is the number of observations), implying the local optimality only for cases when the oracle rate is at least of the order $\sigma n^{1/4}$. In passing, we also get the local optimal results for estimation and posterior contraction problems. Adaptive minimax

results (also for the estimation and posterior contraction problems) over various sparsity classes follow from our local results.

Chapter 3 is based on paper [14]. In this chapter we consider empirical Bayesian inference in the many normal means model in the situation when the high-dimensional mean vector is *multilevel sparse*, that is, most of the entries of the parameter vector are some fixed values. For instance, the traditional *sparse* signal is a particular case (with one level) of multilevel sparse sequences. We apply an empirical Bayes approach, namely we put an appropriate prior modeling the multilevel sparsity and make data-dependent choices of certain parameters of the prior. We establish local (i.e., with rate depending on the "true" parameter) posterior contraction and estimation results. Global adaptive minimax results (for the estimation and posterior contraction problems) over sparsity classes follow from our local results if the sparsity level is of polynomial order. The results are illustrated by simulations.

Chapter 4 is based on paper [16]. In this chapter we study the problem of inference (estimation and *uncertainty quantification* problems) on the unknown parameter in the *biclustering model* by using the penalization method. The underlying biclustering structure is that the high-dimensional parameter consists of a few blocks of equal coordinates. The quality of the inference procedures is measured by the local quantity, the *oracle rate*, which is the best trade-off between the approximation error by a biclustering structure and the complexity of that approximating biclustering structure. The approach is also *robust* in that the additive errors are assumed to satisfy only certain mild condition (allowing non-iid errors with unknown joint distribution). By using the penalization method, we construct a confidence set and establish its local (oracle) optimality. Interestingly, as we demonstrate, there is (almost) *no deceptiveness* issue for the uncertainty quantification problem in the biclustering model. Adaptive minimax results for the biclustering, *stochastic block model* (with implications for network modeling) and *graphon* scales follow from our local results.

Chapter 5 is based on paper [17]. In this chapter we further develop the *general framework of projection structures* introduced in Section 1.5, and study the above mentioned inference problems within this framework by using *empirical Bayes* and *penalization* methods. The main inference problem is *uncertainty quantification*, but on the way we solve the *estimation* and *posterior contraction* problems as well. The approach is *local* in that the quality of the inference procedures is measured by the local quantity, the *oracle rate*, which is the best trade-off between the approximation error by a projection structure and the complexity of that approximating projection structure. The approach is also *robust* in that the stochastic part of the general framework is assumed to satisfy only certain mild condition, the errors may be non-iid with unknown distribution. We introduce the *excessive bias restriction* under which we establish the local (oracle) confidence optimality of the constructed confidence ball. As the proposed general framework unifies a very broad class of high-dimensional models interesting and important on their own right, the obtained general results deliver a whole avenue of results (many new ones and some known in the literature) for particular models as consequences, including *white noise model* and *density estimation* with *smoothness* structure, *linear regression* and *dictionary learning* with *sparsity* structures, *biclustering* and *stochastic block models* with *clustering* structure, *covariance matrix* estimation with *banding* and sparsity struc-

tures, and others. Many adaptive minimax results over various scales follow from our local results.

2

LOCAL ROBUST INFERENCE FOR POSSIBLY SPARSE SEQUENCES

Suppose we observe $X = X^{(\sigma,n)} = (X_1, \ldots, X_n)$:

$$X_i = \theta_i + \sigma\xi_i, \quad i \in [n] = \{1, \ldots, n\}, \tag{2.1}$$

where $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ is an unknown high-dimensional parameter of interest, the ξ_i's are random errors with $\mathbb{E}_\theta \xi_i = 0$, $\mathrm{Var}(\xi_i) \le C_\xi$, $\sigma > 0$ is the known noise intensity. The goal is to make inference about the parameter θ based on the data X: recovery of θ and *uncertainty quantification* by constructing an *optimal confidence set*. We pursue *robust inference* in the sense that the distribution of the error vector $\xi = (\xi_1, \ldots, \xi_n)$ is unknown and can also depend on θ, but assumed to satisfy only certain mild *exchangeable exponential moment condition*; see Condition (B0) in Section 2.1. For inference on θ, we exploit the empirical Bayes approach. We derive non-asymptotic results, which imply asymptotic assertions as well if needed. Possible asymptotic regimes are decreasing noise level $\sigma \to 0$, high-dimensional setup $n \to \infty$ (the leading case for high dimensional models), or their combination, e.g., $\sigma = n^{-1/2}$ and $n \to \infty$.

Useful inference is not possible without some structure on θ. Popular structural assumptions are *smoothness* and *sparsity*, in this chapter we are concerned with the latter. The best studied problem in the sparsity context is that of estimating θ in the many normal means model, a variety of estimation methods and results are available in the literature: [38], [23], [50], [2], [33], [85]. However, even an optimal estimator does not reveal how far it is from θ. It is of importance to quantify this uncertainty, which can be seen as the problem of constructing confidence sets for θ.

Many inference methods have Bayesian connections. For example, even some seemingly non Bayesian estimators can be obtained as certain quantities (like posterior mode for penalized minimum contrast estimators) of the (empirical Bayes) posterior distributions resulting from imposing some specific priors on the parameter; cf. [50] and [2]. Although the Bayesian methodology is used or can be related to in constructing many

(frequentist) inference procedures, only recently the posterior distributions themselves have been studied in the sparsity context: [33], [85], [64], [32], [22], [80], [78].

In this chapter for inference on θ we use an empirical Bayes approach. Since any Bayesian approach always delivers a posterior $\pi(\vartheta|X)$ (in the posteriors for θ, we will use the variable ϑ to distinguish it from the "true" θ), an accompanying problem of interest is the contraction of the resulting (empirical Bayes) posterior to the "true" θ from the frequentist perspective of the "true" measure $\mathbb{P}_\theta$, the distribution of X from (2.1). The quality of posterior is characterized by the posterior contraction rate. We pursue a novel local approach by allowing the posterior contraction rate to be a local quantity, i.e., depending on the true θ, whereas global minimax rates are typically studied in the literature on Bayesian nonparametrics.

A common Bayesian way to model sparsity structure is by the so called two-groups priors. Such a prior puts positive mass on vectors θ with some exact zero coordinates (zero group) and the remaining coordinates (signal group) are drawn from a chosen distribution. So the marginal prior for each coordinate is a mixture of a continuous distribution and a point-mass at zero. In [33] it is shown that for a suitably chosen two-groups prior, the posterior concentrates around the true θ at the minimax rate (as $n \to \infty$) for two sparsity classes, *nearly black vectors* $\ell_0[p_n]$ with p_n nonzero coordinates and *weak* ℓ_q*-balls* $m_q[p_n]$. As pointed out by [33] (also by [50]), the prior distributions of non-zero coordinates should not have too light tails, otherwise one gets sub-optimal rates. The important Gaussian case is for example excluded. This has to do with the so called *over-shrinkage effect* of the normal prior with a fixed mean, which pushes the posterior too much towards the prior mean, missing the true parameter that in general differs from the prior mean. That is why [50] and [33] discard normal priors on non-zero coordinates and use heavy tailed priors. A way to construct such a prior is to put a next level heavy-tailed prior, like half-Cauchy, on the variance in the normal prior, resulting in the so called (one-component) horseshoe prior on θ (cf. [31] and [85]). In the present chapter we show that normal priors are still usable (cf. [64]) and lead to strong local results (even for non-normal models) if combined with empirical Bayes approach.

The main aim in this chapter is to construct confidence sets with optimal properties for possibly sparse sequences. The optimality framework for uncertainty quantification is already discussed in detail in Section 1.4. As we mentioned in Section 1.4, the "deceptiveness" phenomenon is well understood for some smoothness structures (e.g., Sobolev scale), especially in global minimax settings; see [77], [26], [9] and [83]. If we now insist on the optimal size property in (1.14) for all Θ_β, $\beta \in \mathcal{B}$, the coverage relation in (1.14) will not hold for all $\Theta_0 = \Theta_\beta$, but only for $\Theta_0 = \Theta_\beta \backslash \Theta'$, with some set of "deceptive parameters" Θ' removed from Θ_β. In [83] such parameters are called "inconvenient truths" and an explicit construction of a $\theta \in \Theta'$ is given. Examples of non-deceptive parameters are the set of *self-similar* parameters $\Theta_0 = \Theta_{ss}$ introduced by [70] and studied by [25], [26], [83], and the set of *polished tail parameters* $\Theta_0 = \Theta_{pt}$ considered by [83]. In all the above mentioned papers global minimax radial rates (i.e., $r(\theta) = r(\Theta_\beta)$ for all $\theta \in \Theta_\beta$) for specific smoothness structures are studied. A local approach, delivering also the adaptive minimax results for many smoothness structures simultaneously, is considered by [5] for posterior contraction rates and by [9] for constructing optimal confidence balls. In [9], yet a more general (than Θ_{ss} and Θ_{pt}) set of non-deceptive parameters was introduced,

$\Theta_0 = \Theta_{ebr}$, parameters satisfying the so called *excessive bias restriction* (EBR). More on this can be found in Section 2.3.

To the best of our knowledge, there are very few papers about adaptive results on uncertainty quantification (1.14). The case of two nearly black classes is treated by [67], the "general polished tail" condition was introduced in [79] to describe non-deceptive parameters. A restricted scale of nearly black classes is treated in [86], where effectively a version of our EBR condition is used, more on relation to paper [86] can be found in Section 2.5.

In this chapter we introduce a family of normal mixture priors and propose an empirical Bayes procedure (in fact, two procedures). We use the normal likelihood, whereas the true model (2.1) does not have to be normal (and independence of ξ_i's is not required either), but only satisfying some mild Condition (B0) (called *exchangeable exponential moment condition*). There are three distinctive features of our approach: *robust*, *local* and *refined*.

First, *robust* means that our results cover also misspecified models, as we allow the ξ_i's to be not necessarily independent normals (a certain type of error misspecification was also mentioned in a remark of the supplement to [32]), but only satisfying Condition (B0). It turned out that, although we use the normal likelihood (whereas the true model may not be normal) in the Bayesian analysis in the proof of the main results, we can handle the frequentist behavior of the posterior from the perspective of the true measure only on the basis of Condition (B0).

Second, we develop the novel *local* approach, meaning that the radial rate $r(\theta)$ in (1.14) is allowed to be a function of θ, which, in a way, measures the amount of sparsity for each $\theta \in \mathbb{R}^n$: the smaller $r(\theta)$, the more sparse θ. The local radial rate $r(\theta)$ is constructed as the best (smallest) rate over a certain family of local rates, therefore called *oracle rate*. We demonstrate that the local approach is more powerful than global in that we do not need to impose any specific sparsity structure, because the proposed local approach automatically exploits the "effective" sparsity of each underlying θ, and our local results imply a whole panorama of the global minimax results for many scales at once. More on this is in Section 2.2.5.

Third, we derive the local posterior contraction result for the resulting empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ in the *refined non-asymptotic formulation*: $\sup_{\theta\in\mathbb{R}^n} \mathbb{E}_\theta \hat{\pi}(\|\vartheta - \theta\|^2 \geq M_0 r^2(\theta) + M\sigma^2|X) \leq H_0 e^{-m_0 M}$ for some fixed $M_0, H_0, m_0 > 0$ and arbitrary $M \geq 0$, as an exponential non-asymptotic concentration bound in terms of M, uniformly in $\theta \in \mathbb{R}^n$. This formulation provides a rather subtile characterization of the quality of the posterior (finer, than, e.g., asymptotically in terms of the dimension n), allowing subtle analysis for various asymptotic regimes. This result is of interest and importance on its own as it actually establishes the contraction of the empirical Bayes posterior with the local rate $r(\theta)$. Besides, we obtain the oracle estimation result (also in similar refined formulation, finer than traditional oracle inequalities) by constructing an estimator, the empirical Bayes posterior mean, which converges to θ with the local rate $r(\theta)$. This result, besides being an ingredient for the uncertainty quantification problem (1.14), is also of interest and importance on its own as it delivers the same (oracle and minimax) estimation results as in [2] and [50] and posterior convergence results as in [33], obtained for different priors.

Next, we construct a confidence ball by using the empirical Bayes posterior quantities. Since we want the size of our confidence sets to be of an oracle rate order, this comes with the price that the coverage property can hold uniformly only over some set of parameters satisfying the so called *excessive bias restriction* (EBR) $\Theta_0 = \Theta_{\text{eb}} \subseteq \mathbb{R}^n$. The main result consists in establishing the optimality (1.14) of the constructed confidence ball for the optimality framework $\Theta_0 = \Theta_{\text{eb}}$, $\Theta_1 = \mathbb{R}^n$ and the local radial rate $r(\theta)$. The important consequence of our local approach is that a whole panorama of adaptive (global) minimax results (for all the three problems: estimation, posterior contraction rate and confidence sets) over *all* sparsity scales *covered* by $r(\theta)$ (see Section 2.2.5) follow from our local results. In particular, our local results imply the same type of adaptive minimax estimation results over sparsity scales as in [50], and the same type of global minimax results on contraction posterior rates as in [33] (and actually more).

We also treat the situation when $\Theta_0 = \mathbb{R}^n$ in (1.14) by constructing a confidence ball with the radius of the order $\sigma n^{1/4} + r(\theta)$. As we already discussed, the term $\sigma n^{1/4}$ in the size relation is necessary for the uniform coverage to hold. Clearly, this confidence ball will have optimal size only for non-sparse parameters (for which $r(\theta) \geq c\sigma n^{1/4}$).

Although the original motivation of the EBR condition was to remove the deceptive parameters, it turned out to be a very useful notion in the context of uncertainty quantification. In effect, the EBR condition leads to a *new sparsity EBR-scale* which gives the slicing of the entire space that is very suitable for uncertainty quantification. This provides a new perspective at the above mentioned "deceptiveness" issue: basically, each parameter is deceptive (or non deceptive) to some extent. It is the structural parameter of the new EBR-scale that measures the deceptiveness amount, and the (mild and controllable) price for handling deceptive parameters is the effective amount of inflating of the confidence ball that matches the amount of deceptiveness needed to provide a high coverage. The EBR condition and EBR-scale are discussed at length in Section 2.3.

This chapter is organized as follows. In Section 2.1 we introduce the notation and the prior, describe the empirical Bayes procedure, and state the exchangeable exponential moment condition on ξ. In Section 2.2 we introduce the EBR condition and present the main results of this chapter. In Section 2.3 we discuss the EBR condition and the EBR scale at length. Then in Section 2.4 we present a small simulation study. The concluding remarks are provided in Section 2.5. The proofs of the lemmas and theorems are given in Sections 2.6 and 2.7, respectively.

2.1. PRELIMINARIES

First we introduce some notation and a family of normal priors (similar to priors from [9] but geared towards modeling sparsity rather than smoothness). Next, by applying the empirical Bayes approach to the normal likelihood, we derive an empirical Bayes posterior which we will use in the construction of the estimator and the confidence ball. We complete this section with introducing the *exchangeable exponential moment condition* for possibly sparse sequences on the error vector $\xi = (\xi_1, \ldots, \xi_n)$.

2.1.1. NOTATION

Denote the probability measure of X from the model (2.1) by $\mathbb{P}_\theta = \mathbb{P}_\theta^{(\sigma,n)}$, and by $\mathbb{E}_\theta$ the corresponding expectation. For notational simplicity, we often skip the dependence on σ and n. Let $\mathcal{I} = \mathcal{I}_n = \{I : I \subseteq [n]\}$ be the family of all subsets of $[n]$ including the empty set. If the summation range in $\sum_I$ is not specified (for brevity), this means $\sum_{I\in\mathcal{I}}$. Throughout we assume the conventions that $\sum_{i\in\varnothing} a_i = 0$, $\sum_a^b a_i = \sum_{a\le i\le b} a_i$ for any $a_i, a, b \in \mathbb{R}$ and $0\log(c/0) = 0$ (hence $(c/0)^0 = 1$) for any $c > 0$. Let $\theta^2_{(1)} \le \theta^2_{(2)} \le \ldots \le \theta^2_{(n)}$ and $\theta^2_{[1]} \ge \theta^2_{[2]} \ge \ldots \ge \theta^2_{[n]}$ be the ordered values of $\theta_1^2, \ldots, \theta_n^2$. To have some quantity well defined in the sequel, introduce also $0 = \theta^2_{(0)} = \theta^2_{[n+1]}$ and $\theta^2_{[0]} = \theta^2_{(n+1)} = \infty$. If random quantities appear in a relation, this relation should be understood in $\mathbb{P}_\theta$-almost sure sense. Finally, denote $X(I) = (X_i 1\{i \in I\}, i \in [n])$ for $I \in \mathcal{I}$.

2.1.2. MULTIVARIATE NORMAL PRIOR

When deriving all the posterior quantities in the Bayesian analysis below, we will use the normal likelihood $\ell(\theta, X) = (2\pi\sigma^2)^{-n/2}\exp\{-\|X-\theta\|^2/2\sigma^2\}$, which is equivalent to imposing the classical high-dimensional normal model $X = (X_i, i \in \mathbb{N}_n) \sim \bigotimes_{i=1}^n \mathrm{N}(\theta_i, \sigma^2)$. Recall however that the "true" model $X \sim \mathbb{P}_\theta$ is not assumed to be normal, but only satisfying Condition (B0).

To model possible sparsity in the parameter θ, the coordinates of θ can be split into two distinct groups of coordinates of θ: for some $I \in \mathcal{I}$, $\theta_I = (\theta_i, i \in I)$ and $\theta_{I^c} = (\theta_i, i \in I^c)$, so that $\theta = (\theta_I, \theta_{I^c})$. The group of coordinates $\theta_{I^c} = (\theta_i, i \notin I)$ consists of (almost) zeros and $\theta_I = (\theta_i, i \in I)$ is the group of non-zeros coordinates. For any $\theta \in \mathbb{R}^n$ (even "not sparse" one) there is the best (oracle) splitting into two groups, we will come back to this in Section 2.2. To model sparsity, we propose a prior on θ given I as follows:

$$\pi_I = \bigotimes_{i=1}^n \mathrm{N}\big(\mu_i(I), \tau_i^2(I)\big), \quad \mu_i(I) = \mu_i 1\{i \in I\}, \quad \tau_i^2(I) = \sigma^2 K_n(I) 1\{i \in I\}, \tag{2.2}$$

and $K_n(I) = (\frac{en}{|I|} - 1)1\{I \ne \varnothing\}$. The indicators in prior (2.2) ensure the sparsity of the group I^c. The rather specific choice of $K_n(I)$ is made for the sake of concise expressions in later calculations, many other choices are actually possible. By using normal likelihood $\ell(\theta, X) = (2\pi\sigma^2)^{-n/2}\exp\{-\|X-\theta\|^2/2\sigma^2\}$, the corresponding posterior distribution for θ is readily obtained:

$$\pi_I(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}\Big(\frac{\tau_i^2(I)X_i + \sigma^2\mu_i(I)}{\tau_i^2(I)+\sigma^2}, \frac{\tau_i^2(I)\sigma^2}{\tau_i^2(I)+\sigma^2}\Big). \tag{2.3}$$

Next, introduce the prior λ on $\mathcal{I}$, discussed in Section 2.5. For $\kappa > 1$, draw a random set from $\mathcal{I}$ with probabilities

$$\lambda_I = c_{\kappa,n}\exp\big\{-\kappa|I|\log(\tfrac{en}{|I|})\big\} = c_{\kappa,n}(\tfrac{en}{|I|})^{-\kappa|I|}, \quad I \in \mathcal{I}, \tag{2.4}$$

where $c_{\kappa,n}$ is the normalizing constant. Since $(\frac{n}{k})^k \le \binom{n}{k} \le (\frac{en}{k})^k$ and $\binom{n}{0} = 1$,

$$1 = \sum_{I\in\mathcal{I}} \lambda_I = c_{\kappa,n}\sum_{k=0}^n \binom{n}{k}\Big(\frac{en}{k}\Big)^{-\kappa k} \le c_{\kappa,n}\sum_{k=0}^n \Big(\frac{en}{k}\Big)^{-(\kappa-1)k} \le c_{\kappa,n}\sum_{k=0}^n e^{-(\kappa-1)k}, \tag{2.5}$$

so that $c_{\kappa,n} \geq 1 - e^{1-\kappa} > 0$, $n \in \mathbb{N}$. Combining (2.2) and (2.4) gives the mixture prior on θ: $\pi = \sum_{I\in\mathcal{I}} \lambda_I \pi_I$. This leads to the marginal distribution of X: $\mathbb{P}_X = \sum_{I\in\mathcal{I}} \lambda_I \mathbb{P}_{X,I}$, with $\mathbb{P}_{X,I} = \bigotimes_{i=1}^n \mathrm{N}\big(\mu_i(I), \sigma^2 + \tau_i^2(I)\big)$, and the posterior of θ is

$$\pi(\vartheta|X) = \pi_\kappa(\vartheta|X) = \sum_{I\in\mathcal{I}} \pi_I(\vartheta|X)\pi(I|X), \tag{2.6}$$

where $\pi_I(\vartheta|X)$ is defined by (2.3) and the posterior $\pi(I|X)$ for I is

$$\pi(I|X) = \frac{\lambda_I \mathbb{P}_{X,I}}{\mathbb{P}_X} = \frac{\lambda_I \prod_{i=1}^n \phi\big(X_i, \mu_i(I), \sigma^2 + \tau_i^2(I)\big)}{\sum_{J\in\mathcal{I}} \lambda_J \prod_{i=1}^n \phi\big(X_i, \mu_i(J), \sigma^2 + \tau_i^2(J)\big)}. \tag{2.7}$$

2.1.3. EMPIRICAL BAYES POSTERIOR

The parameters μ_i are yet to be chosen in the prior. We choose μ_i by using empirical Bayes approach. The marginal likelihood $\mathbb{P}_X$ is readily maximized with respect to μ_i: $\tilde{\mu}_i = X_i$, which we then substitute instead of μ_i in the expression (2.6) for $\pi(\vartheta|X)$, obtaining the empirical Bayes posterior

$$\tilde{\pi}(\vartheta|X) = \tilde{\pi}_\kappa(\vartheta|X) = \sum_{I\in\mathcal{I}} \tilde{\pi}_I(\vartheta|X)\tilde{\pi}(I|X), \tag{2.8}$$

where the empirical Bayes conditional posterior (recall that $\mathrm{N}(0,0) = \delta_0$)

$$\tilde{\pi}_I(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}\big(X_i 1\{i \in I\}, \tfrac{K_n(I)\sigma^2 1\{i\in I\}}{K_n(I)+1}\big) \tag{2.9}$$

is obtained from (2.3) with $\mu_i = X_i$, and

$$\tilde{\pi}(I|X) = \frac{\lambda_I \mathbb{P}_{X,I}}{\sum_{J\in\mathcal{I}} \lambda_J \mathbb{P}_{X,J}} = \frac{\lambda_I \prod_{i=1}^n \phi(X_i, X_i 1\{i \in I\}, \sigma^2 + \tau_i^2(I))}{\sum_{J\in\mathcal{I}} \lambda_J \prod_{i=1}^n \phi(X_i, X_i 1\{i \in J\}, \sigma^2 + \tau_i^2(J))} \tag{2.10}$$

is the empirical Bayes posterior for $I \in \mathcal{I}$, obtained from (2.7) with $\mu_i(I) = X_i$. Let $\tilde{\mathbb{E}}$ and $\tilde{\mathbb{E}}_I$ be the expectations with respect to the measures $\tilde{\pi}(\vartheta|X)$ and $\tilde{\pi}_I(\vartheta|X)$ respectively. Then $\tilde{\mathbb{E}}_I(\vartheta|X) = X(I) = (X_i 1\{i \in I\}, i \in [n])$. Introduce the *empirical Bayes posterior mean* estimator

$$\tilde{\theta} = \tilde{\mathbb{E}}(\vartheta|X) = \sum_{I\in\mathcal{I}} \tilde{\mathbb{E}}_I(\vartheta|X)\tilde{\pi}(I|X) = \sum_{I\in\mathcal{I}} X(I)\tilde{\pi}(I|X). \tag{2.11}$$

Consider an alternative empirical Bayes posterior. First derive an empirical Bayes variable selector $\hat{I}$ by maximizing $\tilde{\pi}(I|X)$ over $I \in \mathcal{I}$ (any maximizer will do) as follows:

$$\begin{aligned}\hat{I} &= \operatorname*{argmax}_{I\in\mathcal{I}} \tilde{\pi}(I|X) = \operatorname*{argmax}_{I\in\mathcal{I}} \lambda_I \mathbb{P}_{X,I} = \operatorname*{argmax}_{I\in\mathcal{I}} \Big\{ -\sum_{i\in I^c} \tfrac{X_i^2}{2\sigma^2} - \tfrac{|I|}{2}\log(K_n(I)+1) + \log\lambda_I \Big\} \\ &= \operatorname*{argmin}_{I\in\mathcal{I}} \Big\{ \sum_{i\in I^c} X_i^2 + (2\kappa+1)\sigma^2 |I| \log\big(\tfrac{en}{|I|}\big) \Big\},\end{aligned} \tag{2.12}$$

which is reminiscent of the penalization procedure from [23] (cf. also [2]). Now plugging in $\hat{I}$ into $\tilde{\pi}_I(\vartheta|X)$ defined by (2.9) yields another empirical (now with respect to μ_i's and I) Bayes posterior and the corresponding empirical Bayes mean estimator for θ:

$$\check{\pi}(\vartheta|X) = \tilde{\pi}_{\hat{I}}(\vartheta|X), \quad \check{\theta} = \check{\mathbb{E}}(\vartheta|X) = X(\hat{I}) = (X_i 1\{i \in \hat{I}\}, i \in [n]), \tag{2.13}$$

where $\check{\mathbb{E}}$ denotes the expectation with respect to the measure $\check{\pi}(\vartheta|X)$.

2.1.4. EXCHANGEABLE EXPONENTIAL MOMENT CONDITION ON THE ERRORS

Condition (A1) (called *exchangeable exponential moment condition*, defined in Section 1.5) on the error vector $\xi = (\xi_1, \dots, \xi_n)$ for possibly sparse sequences can be formulated as follows:

CONDITION (B0). The random variables ξ_i's from (2.1) satisfy: $\mathbb{E}_\theta \xi_i = 0$, $\mathrm{Var}(\xi_i) \le C_\xi$, $i \in [n]$; and for some $\alpha > 0$ (without loss of generality assume $C_\xi = 1$ and $\alpha \in (0,1]$),

$$\mathbb{E}_\theta \exp\Big\{\alpha \sum_{i\in I} \xi_i^2\Big\} \le e^{|I|} \quad \text{for all } I \in \mathcal{I},\ \theta \in \mathbb{R}^n. \tag{B0}$$

There is no need to assume $\mathrm{Var}(\xi_i) \le C_\xi$ as this follows from (B0), but we provide this just for reader's convenience. In case of independent normal errors, some bounds in the proofs can be sharpened; possible refinements are discussed in Section 2.5.

As we mentioned in Chapter 1, the sub-gaussianity condition (1.20) and Condition (A1) (in particular, Condition (B0)) are close, but in general incomparable. For example, let $\xi_i = \xi_0$, $i \in [n]$, for some bounded random variable ξ_0 (say, uniform on $[-1,1]$), then Condition (B0) trivially holds whereas the sub-gaussianity condition is not fulfilled. It is easy to see that the sub-gaussianity condition is equivalent to Condition (B0) for independent ξ_i's.

Moreover, the ξ_i's do not have to be normal and do not have to be even independent. Suppose that the ξ_i's follow an autoregressive model AR(1) with normal white noise:

$$\xi_k = \gamma \xi_{k-1} + \epsilon_k, \quad \epsilon_k \overset{\text{ind}}{\sim} \mathrm{N}(0,1), \quad k \in [n]; \quad \xi_0 = 0, \quad |\gamma| < 1. \tag{2.14}$$

Let us show that Condition (B0) is fulfilled for the vector $\xi = (\xi_i, i \in [n])$. We have that for any $k > l$, $\xi_k = \gamma^{k-l}\xi_l + \gamma^{k-l-1}\epsilon_{l+1} + \dots + \epsilon_k = \gamma^{k-l}\xi_l + Z_{k-l}$, where $Z_{k'} \sim \mathrm{N}(0, \sigma^2_{k'})$ with $\sigma^2_{k'} = 1 + \gamma^2 + \dots + \gamma^{2(k'-1)} \le \frac{1}{1-\gamma^2} \triangleq \sigma_0^2$. Clearly, for any $I \in \mathcal{I}$, there are $1 \le k_1 < k_2 < \dots < k_{|I|} \le n$ such that $\sum_{i\in I} \xi_i^2 = \sum_{i=1}^{|I|} \xi^2_{k_i}$. Denote $\mathcal{F}_m = \sigma(\xi_{k_i}, 1 \le i \le m)$, $m \in [|I|]$, the σ-algebra generated by $\{\xi_{k_i}, 1 \le i \le m\}$. Choose α and γ in such a way that $0 < \frac{2\alpha\gamma^2}{1-4\alpha\sigma_0^2} \le \alpha$. By using the elementary identity (S1) (from Section 2.5), we first evaluate the conditional expectation

$$\begin{aligned}
&\mathbb{E}_\theta\big(e^{\alpha(\xi^2_{k_{m-1}} + \xi^2_{k_m})} | \mathcal{F}_{m-1}\big) \le e^{\alpha\xi^2_{k_{m-1}}}\, \mathbb{E}_\theta\big(e^{2\alpha\xi^2_{k_m}} | \mathcal{F}_{m-1}\big) \\
&= \exp\Big\{\big(\alpha + \tfrac{2\alpha\gamma^{2(k_m - k_{m-1})}}{1-4\alpha\sigma^2_{k_m,k_{m-1}}}\big)\xi^2_{k_{m-1}} - \tfrac{1}{2}\log(1 - 4\alpha\sigma^2_{k_m,k_{m-1}})\Big\} \le (1-4\alpha\sigma_0^2)^{-1/2} e^{2\alpha\xi^2_{k_{m-1}}}.
\end{aligned}$$

Iterating the above conditional expectation argument, we establish Condition (B0) for the sequence (2.14):

$$\begin{aligned}
\mathbb{E}_\theta \exp\Big\{\alpha \sum_{i\in I} \xi_i^2\Big\} &= \mathbb{E}_\theta \mathbb{E}_\theta\Big[\exp\Big\{\alpha \sum_{i\in I} \xi_i^2\Big\}\Big| \mathcal{F}_{|I|-1}\Big] = (1-4\alpha\sigma_0^2)^{-1/2} \mathbb{E}_\theta\Big[\exp\Big\{\alpha \sum_{i=1}^{|I|-2} \xi^2_{k_i}\Big\} e^{2\alpha\xi^2_{k_{|I|-1}}}\Big] \\
&\le \dots \le (1-4\alpha\sigma_0^2)^{-|I|/2} = e^{B|I|}, \qquad \text{with} \quad B = \log(1-4\alpha\sigma_0^2)^{-1/2}.
\end{aligned}$$

Then Condition (B0) holds in view of Remark 1.4.

In the proof of Theorem 2.1 below, we will need a bound for $\mathbb{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2$, $I\in\mathcal{I}$. Actually, Condition (B0) ensures such a bound. Indeed, since $x^2\le e^{2x}$ for all $x\ge 0$, by using the Hölder inequality and (B0), we derive that for any $t\in(0,\alpha]$

$$\mathbb{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2=\frac{4}{t^2}\mathbb{E}_\theta\big(\frac{t}{2}\sum_{i\in I}\xi_i^2\big)^2\le\frac{4}{t^2}\mathbb{E}_\theta e^{t\sum_{i\in I}\xi_i^2}\le\frac{4}{t^2}\big[\mathbb{E}_\theta e^{\alpha\sum_{i\in I}\xi_i^2}\big]^{t/\alpha}\le\frac{4}{t^2}e^{t\alpha^{-1}|I|}.$$

Thus Condition (B0) implies that for any $\rho\in(0,1/2]$ and any $I\in\mathcal{I}$,

$$\mathbb{E}_\theta\big(\textstyle\sum_{i\in I}\xi_i^2\big)^2\le\frac{1}{(\alpha\rho)^2}e^{2\rho|I|}. \tag{2.15}$$

2.2. MAIN RESULTS

In this section we give the main results of this chapter. From now on, by $\hat\pi(\vartheta|X)$ (with corresponding expectation $\hat{\mathbb{E}}(\cdot|X)$) we denote either $\tilde\pi(\vartheta|X)$ defined by (2.8) or $\check\pi(\vartheta|X)$ defined by (2.13), and $\hat\theta$ will stand either for $\tilde\theta$ defined by (2.11) or for $\check\theta$ defined by (2.13). Also, $\hat\pi(I\in\mathcal{G}|X)$ should be read as $\tilde\pi(I\in\mathcal{G}|X)$ in case $\hat\pi=\tilde\pi$, and as $\mathbb{1}\{\hat I\in\mathcal{G}\}$ in case $\hat\pi=\check\pi$, for all $\mathcal{G}\subseteq\mathcal{I}$ that appear in what follows. Hence, $\hat\pi(I|X)=\tilde\pi(I|X)$ and $\mathbb{E}_\theta\hat\pi(I\in\mathcal{G}|X)=\mathbb{E}_\theta\tilde\pi(I\in\mathcal{G}|X)$ in the former case, and $\hat\pi(I|X)=\mathbb{1}\{\hat I=I\}$ and $\mathbb{E}_\theta\hat\pi(I\in\mathcal{G}|X)=\mathbb{P}_\theta(\hat I\in\mathcal{G})$ in the latter case.

2.2.1. ORACLE RATE

The empirical Bayes posterior $\hat\pi(\vartheta|X)$ is a random mixture over $\tilde\pi_I(\vartheta|X)$, $I\in\mathcal{I}$. From the $\mathbb{P}_\theta$-perspective, each posterior $\tilde\pi_I(\vartheta|X)$ (and the corresponding estimator $\tilde{\mathbb{E}}_I(\vartheta|X)=X(I)$) contracts to the true parameter θ with the local rate $R^2(I,\theta)=\sum_{i\in I^c}\theta_i^2+\sigma^2|I|$. Indeed, since $\tilde{\mathbb{E}}_I(\vartheta|X)=X(I)=(X_i1\{i\in I\},i\in\mathbb{N}_n)$, (2.9) and the Markov inequality yield

$$\mathbb{E}_\theta\tilde\pi_I(\|\vartheta-\theta\|^2\ge M^2R^2(I,\theta)|X)\le\frac{\mathbb{E}_\theta\|X(I)-\theta\|^2+\frac{K_n(I)\sigma^2|I|}{K_n(I)+1}}{M^2R^2(I,\theta)}\le\frac{2}{M^2}.$$

For each $\theta\in\mathbb{R}^n$, among $I\in\mathcal{I}$ there exists the best choice $I_o=I_o(\theta)=I_o(\theta,\sigma)$ (called the *R-oracle*) corresponding to the fastest local rate $R^2(\theta)=R^2(\theta,\mathcal{I})=\min_{I\in\mathcal{I}}R^2(I,\theta)=\sum_{i\in I_o^c}\theta_i^2+\sigma^2|I_o|$. Ideally, we would like to *mimic* the R-oracle, i.e., to construct an empirical Bayesian procedure (e.g., $\hat\pi(\vartheta|X)$) which performs as good as the oracle empirical Bayes posterior $\tilde\pi_{I_o}(\vartheta|X)$ without knowing I_o, uniformly in $\theta\in\mathbb{R}^n$. However, the lower bounds for the estimation problem (hence, also for the posterior contraction problem), obtained by [37] and later by [23], show that it is impossible to mimic the R-oracle and a logarithmic factor is the unavoidable price for the uniformity over $\mathbb{R}^n$ (otherwise this would contradict to the minimax lower bound over the scale of sparsity classes, cf. [23]). Therefore only a modification of the risk R-oracle where the variance term $\sigma^2|I_o|$ is inflated with the factor $\log(en/|I_o|)$ (thought of as payment for not knowing I_o) is "mimicable".

The above discussion motivates the following definition. Introduce the family of local rates

$$r^2(I,\theta)=r^2_\sigma(I,\theta)=\sum_{i\in I^c}\theta_i^2+\sigma^2|I|\log(\tfrac{en}{|I|})=B(I,\theta)+V(I),\quad I\in\mathcal{I}, \tag{2.16}$$

where $B(I,\theta) = \sum_{i\in I^c}\theta_i^2$ is the bias part of the rate and $V(I) = V(I,\sigma,n) = \sigma^2|I|\log(\frac{en}{|I|})$ is the adjusted variance part, the variance term $\sigma^2|I|$ of the rate $R(I,\theta)$ multiplied by the logarithmic factor $\log(\frac{en}{|I|})$. There exists the best choice $I_o = I_o(\theta) = I_o(\theta,\sigma^2) = I_o(\theta,\sigma^2,n) \in \mathcal{I}$ (called *oracle*) at which the rate (2.16) is minimal:

$$r^2(\theta) = r^2(\theta,\mathcal{I}) = \min_{I\in\mathcal{I}} r^2(I,\theta) = r^2(I_o(\theta),\theta) = B(I_o(\theta),\theta) + V(I_o(\theta)), \tag{2.17}$$

called the *oracle rate*. Note that the oracle I_o may not be unique (but $|I_o|$ is unique) as some coordinates of θ can coincide, in that case take the one with the earliest coordinates. Clearly, $I_o = \{i \in [n] : \theta_i^2 \ge \theta_{[i_o]}^2\}$, where $i_o = |I_o| = \operatorname{argmin}_{k\in[n]_0}\{\sum_{i=1}^{n-k}\theta_{(i)}^2 + \sigma^2 k\log(\frac{en}{k})\}$. Thus the oracle I_o classifies the coordinates $(\theta_i, i \in I_o)$ as *significant* and the coordinates $(\theta_i, i \in I_o^c)$ as *insignificant*. The bias related term $B(I_o(\theta),\theta) = \sum_{i\in I_o^c}\theta_i^2 = \sum_{i=1}^{n-i_o}\theta_{(i)}^2$ of the oracle rate is called the *excessive bias*. This is the error the oracle makes when setting insignificant coordinates of θ to zero. The variance related term $\sigma^2|I_o|\log(\frac{en}{|I_o|})$ is the error the oracle makes when recovering the significant coordinates (the log factor is the payment for not knowing the locations). The definition (2.17) of the oracle I_o implies a certain characterization of the significant (and insignificant) coordinates $\{\theta_{[i]}, i = 1,\dots,i_o\}$, which is provided in Section 2.5.

Introduce a family of the so called τ-oracles $I_o^\tau = I_o^\tau(\theta) = I_o(\theta,\tau\sigma^2)$, $\tau \ge 0$ and let $i_\tau = |I_o^\tau(\theta)|$ be the corresponding cardinalities. A τ-oracle $I_o^\tau(\theta)$ is just the usual oracle defined by (2.17) with σ^2 substituted by $\tau\sigma^2$, the oracle itself is the τ-oracle with $\tau = 1$: $I_o(\theta) = I_o^1(\theta)$. Notice that $I_o^{\tau_1} \subseteq I_o^{\tau_2}$ if $\tau_1 \ge \tau_2$. For $\tau \downarrow 0$, $r^2(\theta, I_o^\tau) \downarrow 0$ and the "limiting" τ-oracle recovers the active index set $I^* = I^*(\theta) = \{i \in [n] : \theta_i \ne 0\}$ in the sense that $I_o^\tau \uparrow I^*$ as $\tau \downarrow 0$. Informally, since the τ-oracle is defined by substituting $\tau\sigma^2$ instead of σ^2 in the oracle rate, one can think of the τ-oracle with $\tau \in [0,1)$ as if X is observed with a "magnifying glass" since the error variance is reduced by the factor τ so that the τ-oracle can distinguish more coordinates from zero. In the case $\tau > 1$, the error variance in (2.1) increases by the factor τ (as if the observations X_i's get blurred), resulting in a smaller set of significant coordinates recovered by the τ-oracle. However, all τ-oracle rates $r^2(\theta, I_o^\tau)$, $\tau > 0$, are related to the oracle rate $r^2(\theta) = r^2(\theta, I_o^1)$ by the trivial relations

$$r^2(\theta) \le r^2(\theta, I_o^\tau) \le \tau r^2(\theta) \quad \text{for } \tau \ge 1, \quad r^2(\theta, I_o^\tau) \le r^2(\theta) \le \tau^{-1} r^2(\theta, I_o^\tau) \quad \text{for } 0 < \tau < 1.$$

So, in principle we can obtain the result for any τ-oracle rate $r^2(\theta, I_o^\tau)$ via the result for the oracle rate $r^2(\theta)$ and vice versa, but at the price of some multiplicative constant.

Actually, we can look at all τ-oracles I_o^τ, $\tau \ge 0$, from the following general perspective. Introduce a family of $n+1$ sets

$$\mathcal{I}_o = \{I_o(k), k \in [n]_0\}, \quad \text{where} \quad I_o(k) = \{i \in [n] : \theta_i^2 \ge \theta_{[k]}^2\}. \tag{2.18}$$

Clearly, these are embedded sets $\varnothing \triangleq I_o(0) \subseteq I_o(1) \subseteq I_o(2) \subseteq \dots \subseteq I_o(n) = [n]$. Now notice that the oracle set $I_o(\theta)$ and actually all τ-oracles $I_o^\tau(\theta)$, $\tau \ge 0$, are all from this family, in fact, $I_o = I_o(i_o)$ and $I_o^\tau = I_o(i_\tau)$.

2.2.2. CONTRACTION RESULTS WITH ORACLE RATE

First, introduce a technical condition on the parameter κ appearing in (2.4).

CONDITION (B1). The parameter κ of the prior λ defined by (2.4) satisfies

$$\kappa > \bar{\kappa} \triangleq (16-\alpha)/(4\alpha), \tag{B1}$$

where $\alpha > 0$ is from Condition (B0).

The following theorem establishes that the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ (which is either $\tilde{\pi}(\vartheta|X)$ defined by (2.8) or $\check{\pi}(\vartheta|X)$ defined by (2.13)) contracts to θ with the oracle rate $r(\theta)$ from the frequentist $\mathbb{P}_\theta$-perspective, and the empirical Bayes posterior mean $\hat{\theta}$ (which is either $\tilde{\theta}$ defined by (2.11) or $\check{\theta}$ defined by (2.13)) converges to θ with the oracle rate $r(\theta)$, uniformly over the entire parameter space.

Theorem 2.1. *Let Conditions (B0) and (B1) be fulfilled. Then there exist positive constants $M_0, M_1, H_0, H_1, m_0, m_1$ such that for any $\theta \in \mathbb{R}^n$ and any $M \geq 0$,*

$$\mathbb{E}_\theta \hat{\pi}\big(\|\vartheta-\theta\|^2 \geq M_0 r^2(\theta) + M\sigma^2|X\big) \leq H_0 e^{-m_0 M}, \tag{i}$$

$$\mathbb{P}_\theta\big(\|\hat{\theta}-\theta\|^2 \geq M_1 r^2(\theta) + M\sigma^2\big) \leq H_1 e^{-m_1 M}. \tag{ii}$$

Remark 2.1. *Notice that already claim (i) of the above theorem contains an oracle bound for the estimator $\hat{\theta}$. Indeed, by Jensen's inequality, we derive the oracle inequality*

$$\begin{aligned}\mathbb{E}_\theta\|\hat{\theta}-\theta\|^2 \leq \mathbb{E}_\theta\hat{\mathbb{E}}(\|\vartheta-\theta\|^2|X) &\leq M_0 r^2(\theta) + H_0\int_0^{+\infty} e^{-m_0 u/\sigma^2}\,du \\ &= M_0 r^2(\theta) + \frac{H_0\sigma^2}{m_0}.\end{aligned} \tag{2.19}$$

Similarly we can show that also (ii) implies (2.19). This means that claim (ii) is actually stronger (and more refined) than (2.19) and therefore requires a separate proof.

Remark 2.2. *A few more remarks on the theorem are in order.*

(i) *The above local result implies the minimax optimality over various sparsity scales, see Section 2.2.5 for more detail on this.*

(ii) *The constants $M_0, M_1, H_0, H_1, m_0, m_1 > 0$ in the theorem depend only on α and some also on κ, the exact expressions can be found in the proof.*

(iii) *The non-asymptotic exponential bounds in terms of the constant M from the expression $M' r^2(\theta) + M\sigma^2$ (with some fixed M') in claims (i) and (ii) of the theorem provide a very refined characterization of the quality of the posterior $\hat{\pi}(\vartheta|X)$ and estimator $\hat{\theta}$, finer than, e.g., the traditional oracle inequalities like (2.19). This refined formulation allows for subtle analysis in various asymptotic regimes ($n \to \infty$, $\sigma \to 0$, or their combination) as we can let M depend in any way on n, σ, or both.*

One more technical definition will be used. For constant α from Condition (B0), define

$$\bar{\tau} = \bar{\tau}(\kappa, \alpha) \triangleq 3(\kappa\alpha + \tfrac{\alpha}{2} + 1)/\alpha. \tag{2.20}$$

The next theorem describes the frequentist behavior of the selector $\hat{I}$ and the empirical Bayes posterior for I, saying basically that $\hat{I}$ and $\hat{\pi}(I|X)$ "live" on a certain set that

is, in a sense, almost as good as the oracle $I_o = I_o(\theta)$ defined by (2.17). For any $\theta \in \mathbb{R}^n$, introduce

$$I_* = I_*(\theta) \triangleq I_o^{\tau_0}(\theta) = I_o(\theta, \tau_0\sigma^2), \quad i_* = |I_*|, \tag{2.21}$$

where we fix some $\varrho \in (0,1)$ and $\tau_0 > \frac{1+\varrho}{1-\varrho}\bar{\tau}$, $\bar{\tau}$ is defined by (2.20). For example, we can take $\varrho = 0.1$ and $\tau_0 = \frac{11}{9}\bar{\tau} + 0.1$.

Theorem 2.2. *Let Condition (B0) be fulfilled. The following relations hold for any $\theta \in \mathbb{R}^n$, $M \geq 0$.*

(i) Let $\kappa > \frac{6+\alpha}{2\alpha}$ (Condition (B1) implies this). There exist $M_0', H_0' > 0$ such that

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : |I| \log(\tfrac{en}{|I|}) \geq M_0' |I \cap I_o| \log(\tfrac{en}{|I \cap I_o|}) + M \big| X\big) \leq H_0' e^{-M}.$$

(ii) Let $\kappa > \alpha^{-1} - \frac{1}{2}$ (Condition (B1) implies this), $\bar{\tau}$ be defined by (2.20). Fix any $I' \in \mathcal{I}$. Then there exist $H_1', m_0' > 0$ (independent of θ and I') such that

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : \sum_{i \in I' \setminus I} \frac{\theta_i^2}{\sigma^2} \geq \bar{\tau} |I \cup I'| \log(\tfrac{en}{|I \cup I'|}) + M |X\big) \leq H_1' e^{-m_0' M}.$$

In particular, let I_* be defined by (2.21), then there exist $\alpha_1', m_1' > 0$ such that

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : |I| \log(\tfrac{en}{|I|}) \leq \varrho |I_*| \log(\tfrac{en}{|I_*|}) - M \big| X\big) \leq H_1' \big(\tfrac{en}{|I_*|}\big)^{-\alpha_1' |I_*|} e^{-m_1' M}. \tag{2.22}$$

(iii) Let Condition (B1) be fulfilled, c_1, c_2, c_3 be the constants defined in Lemma 2.2. Then

$$\mathbb{E}_\theta \hat{\pi}(I \in \mathcal{I} : r^2(I, \theta) \geq c_3 r^2(\theta) + M\sigma^2 | X) \leq C_0 e^{-c_2 M}, \quad \text{where } C_0 = (1 - e^{1-c_1})^{-1}.$$

Remark 2.3. *The assertion (2.22) holds also for I_* defined differently:*

$$I_* = I_*(\theta) = I_*(\theta, \tau) = I_*(\theta, \tau, \varrho) = \{i \in [n] : \theta_i^2 \geq \theta_{[i_*]}^2\} \tag{2.23}$$

and $i_* = i_*(\tau, \varrho, \theta) = \max\{k \in [n]_0 : \sum_{\varrho k}^k \theta_{[i]}^2 \geq \tau(1-\varrho)\sigma^2 k \log(\frac{en}{k})\}$ for some $\tau \geq \tau_0$. Indeed, the only difference in the proof of (2.22) for I_* defined by (2.23) is that, instead of (2.40), we have the bound

$$\sum_{i \in I_* \setminus I} \frac{\theta_i^2}{\sigma^2} \geq \sum_{\varrho |I_*|}^{|I_*|} \frac{\theta_{[i]}^2}{\sigma^2} \geq \tau(1-\varrho)|I_*| \log(\tfrac{en}{|I_*|}) \geq \tau' |I \cup I_*| \log(\tfrac{en}{|I \cup I_*|}) + \tau' M,$$

so that $m_1' = \tau' m_0'$ in this case and the rest of the proof is the same.

For any $\theta \in \mathbb{R}^n$, the set $I_*(\theta)$ is the representative from the family $\mathcal{I}_o$ (defined by (2.18)) that consists of "distinctly significant" coordinates of θ such that $\hat{\pi}(I|X)$ makes (almost) no mistake for selecting a big proportion of this set. Recall that for $\tau_1 \geq \tau_2$, $I_o(i_{\tau_1}) = I_o^{\tau_1} \subseteq I_o^{\tau_2} = I_o(i_{\tau_2})$ so that $i_{\tau_1} \leq i_{\tau_2}$. Since $I_* = I_o^{\tau_0}$ for $\tau_0 > 1$, we have $I_* \subseteq I_o$. So, the claims of the above theorem roughly mean that $\hat{\pi}(I|X)$ (i.e., the selector $\hat{I}$ and the posterior $\tilde{\pi}(I|X)$) lives in the "shell" $\{I : I_o(K^{-1} i_*) \subseteq I \subseteq I_o(K i_o)\}$ for some sufficiently

large K, where the sets $I_o(k)$ are defined by (2.18). So, if I_* and I_o are "close" to each other (i.e., $i_o \le Ci_*$ for some $C > 0$), then $\hat{\pi}(I|X)$ recovers well the oracle structure I_o. The case $i_* \ll i_o$ is problematic (the corresponding θ is "deceptive") as the living shell for $\hat{\pi}(I|X)$ is then too wide. Property (i) of Theorem 2.2 claims good "over-dimensionality" control of $\hat{\pi}(I|X)$ in terms of the oracle I_o. In other words, the method does a good job in assigning insignificant coordinates to zeros, for any $\theta \in \mathbb{R}^n$. On the other hand, there is no full "under-dimensionality" control for $\hat{\pi}(I|X)$, as property (ii) is only in terms of the set I_* (which may be much "smaller" than I_o): basically for deceptive θ's, $\hat{\pi}(I|X)$ can make relatively many errors by assigning many significantly non-zero coordinates to zeros.

This is reminiscent of the same asymmetric situation for adaption to smoothness where it is also possible to control under-smoothing (e.g., by penalization procedures or Lepski's method), but not over-smoothing. In view of the lower bound results mentioned in the introduction of this thesis, this is not an artefact of the method, it is a fundamental, unavoidable problem. It occurs for the so called deceptive parameters θ that have many smallish coordinates, just slightly under the noise level. Interestingly, controlling over-dimensionality (or under-smoothing for smoothness structures) is enough for solving adaptive estimation problem, but not for uncertainty quantification where we need both over-dimensionality control (for the optimal size) and under-dimensionality control (for the coverage). This is possible for the non-deceptive parameters described by the so called EBR condition and introduced in the next section.

2.2.3. CONFIDENCE BALL UNDER EXCESSIVE BIAS RESTRICTION

Theorem 2.1 establishes the strong local optimal properties of the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ and the empirical Bayes posterior mean $\hat{\theta}$, but these do not solve the uncertainty quantification problem yet. Let us construct a confidence ball by using the empirical Bayes posterior $\check{\pi}(\vartheta|X)$ defined by (2.13). Since $\check{\pi}(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}(\check{\theta}_i, \check{\sigma}_i^2)$ with $\check{\theta}_i = X_i \mathbb{1}\{i \in \hat{I}\}$ and $\check{\sigma}_i^2 = (1 - |\hat{I}|/en)\sigma^2 \mathbb{1}\{i \in \hat{I}\}$, denoting by $\chi^2_{k,\alpha}$ the $(1-\alpha)$-quantile of χ^2_k-distribution we have

$$\check{\pi}\big(\|\vartheta - \check{\theta}\|^2 \le \sigma^2 \chi^2_{|\hat{I}|,\alpha} | X\big) \ge \check{\pi}\big(\|\vartheta - \check{\theta}\|^2 \le (1 - |\hat{I}|/en)\sigma^2 \chi^2_{|\hat{I}|,\alpha} | X\big) = 1 - \alpha.$$

But $\chi^2_{|\hat{I}|,\alpha}$ is bounded by a constant multiple of $|\hat{I}|$, and hence for simplicity the latter can replace the former to obtain a credible ball. This leads to $B(\check{\theta}, M\sigma|\hat{I}|^{1/2})$ as a credible ball for θ, which can be guaranteed to have at least a given level of credibility by choosing a sufficiently large constant M. From (i) of Theorem 2.2 it follows that $|\hat{I}|$ is of the order $|I_o|$. However, it is clear that $B(\check{\theta}, M\sigma|\hat{I}|^{1/2})$ cannot have a guaranteed coverage, since otherwise the center $\check{\theta}$ would be an estimator that mimics the R-oracle uniformly in $\theta \in \mathbb{R}^n$, which is impossible as we discussed earlier. Hence to obtain coverage, the order of the radius of any confidence ball must contain a logarithmic factor. This leads us to the inflated credible ball $B(\check{\theta}, M\hat{r})$, where

$$\hat{r}^2 = \hat{r}^2(X) = \sigma^2 + \sigma^2 |\hat{I}| \log(en/|\hat{I}|). \tag{2.24}$$

The empirical Bayes posterior $\check{\pi}(\vartheta|X)$ is well concentrated (in fact, in a ball of the size $M\sigma^2|I_o|$), but not around the truth, rather around its mean $\check{\theta}$ which in general is away

from the truth by the distance at most of the order of the oracle rate $r(\theta)$. We can also construct a confidence ball by using the posterior $\tilde{\pi}(\vartheta|X)$ defined by (2.8) with the same resulting properties, but with more involved mathematical derivations. Property (i) of Theorem 2.2 means that $\hat{r}^2$ is at most of the order of the variance part of the oracle rate $r^2(\theta)$, so the size property holds uniformly over $\theta \in \mathbb{R}^n$. But this goes at the expense of the coverage, namely, the coverage property does not hold uniformly.

Indeed, according to Theorem 2.2, $\varrho\sigma^2|I_*|\log(en/|I_*|) - M\sigma^2 \le \hat{r}^2 \le M_0'\sigma^2|I_o|\log(en/|I_o|) + M\sigma^2$ with large probability. But this shell can be wide if $\sigma^2|I_*|\log(en/|I_*|) \ll \sigma^2|I_o|\log(en/|I_o|)$. If this happens (for deceptive θ's), then the coverage property of the ball $B(\check{\theta}, M\hat{r})$ cannot be guaranteed because its radius can be of a smaller order than the oracle rate $r^2(\theta)$. This problem will not occur for those (non-deceptive) θ's for which the bias part of the rate $r^2(I_*,\theta)$ (see definition (2.16)) is within a multiple of its variance part $\sigma^2|I_*|\log(en/|I_*|)$. Indeed, then $\sigma^2|I_*|\log(en/|I_*|)$ must be at least some multiple of $r^2(I_*,\theta)$ which is in turn bigger than the oracle rate $r^2(\theta)$ by the definition (2.17) of the oracle. This means that $\sigma^2|I_*|\log(en/|I_*|)$ is at least of the oracle rate order, which, together with (ii) of Theorem 2.2, imply that $\hat{r}$ is also at least of the oracle rate order, resulting in a good coverage of the confidence ball $B(\check{\theta}, M_2\hat{r} + M\sigma^2)$ for some M_2 and sufficiently large M. This discussion motivates introducing the following condition.

CONDITION EBR. We say that $\theta \in \mathbb{R}^n$ satisfies the *excessive bias restriction* (EBR) condition with structural parameter $t \ge 0$ if $\theta \in \Theta_{\mathrm{eb}}(t)$, where the corresponding set (called the *EBR class*) is

$$\Theta_{\mathrm{eb}}(t) = \Theta_{\mathrm{eb}}(t,\tau_0) = \Big\{\theta \in \mathbb{R}^n : \textstyle\sum_{i\in I_*^c}\theta_i^2 \le t\sigma^2\big(1 + |I_*|\log(\frac{en}{|I_*|})\big)\Big\}, \tag{2.25}$$

where the set $I_* = I_o^{\tau_0}$ is defined by (2.21).

The condition EBR essentially requires that the bias part of the rate $r^2(I_*,\theta)$ is dominated by a multiple of its variance part (additional σ^2 is needed to handle the case $I_* = \varnothing$). This is obviously satisfied also for the rate $r^2(I',\theta)$ for all $I' \in \mathcal{I}_o$ such that $I_* \subseteq I'$ (hence also for the oracle I_o since $I_* \subseteq I_o$), where the family $\mathcal{I}_o$ is defined by (2.18). Besides, $\Theta_{\mathrm{eb}}(t_1) \subseteq \Theta_{\mathrm{eb}}(t_2)$ for $t_1 \le t_2$, and, by the definition of I_*, $\mathbb{R}^n = \Theta_{\mathrm{eb}}(\tau_0 n)$.

The EBR condition introduced in [86] is essentially a version of our EBR condition adopted the sparsity scale within the grand space $\ell_0[p_n]$ with $p_n = o(n)$, under the asymptotic regime $n \to \infty$. The EBR condition (and relation to the EBR from [86]) is discussed in more detail in Sections 2.3 and 2.5.

Now we can use the center $\hat{\theta}$ and the radius $\hat{r}$ in constructing a confidence ball for θ. The following theorem, which is the main result in this chapter, describes the coverage and size properties of the confidence ball based on $\hat{\theta}$ and $\hat{r}$.

Theorem 2.3. *Let Conditions* (B0) *and* (B1) *be fulfilled. Then there exist constants* $M_2, H_2, m_2 > 0$ *such that for any* $t, M \ge 0$, *and with* $\hat{R}_M^2 = \hat{R}_M^2(M_2) = (t+1)M_2\hat{r}^2 + (t+2)M\sigma^2$,

$$\sup_{\theta\in\Theta_{\mathrm{eb}}(t)} \mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, \hat{R}_M)\big) \le H_2 e^{-m_2 M},$$
$$\sup_{\theta\in\mathbb{R}^n} \mathbb{P}_\theta\big(\hat{r}^2 \ge M_0'\sigma^2|I_o|\log(\tfrac{en}{|I_o|}) + (M+1)\sigma^2\big) \le H_0' e^{-M},$$

where $\Theta_{\mathrm{eb}}(t)$ is defined by (2.25), the constants M_0', H_0' are defined in Theorem 2.2.

Remark 2.4. *Let the quantity $b(\theta)$ (called* excessive bias ratio*) be defined by*

$$b(\theta) = b(\theta, \tau_0) = \frac{\sum_{i \in I_*^c} \theta_i^2}{\sigma^2 + \sigma^2 |I_*| \log(en/|I_*|)} = \frac{\sum_{i=i_*+1}^{n} \theta_{[i]}^2}{\sigma^2 + \sigma^2 i_* \log(en/i_*)} = \frac{B(I_*, \theta)}{\sigma^2 + V(I_*, \theta)}. \quad (2.26)$$

Note that, when proving Theorem 2.3, we actually established the following local assertions: there exist constants $M_2, \alpha_1, m_1'', H_2, m_2 > 0$ such that for any $\theta \in \mathbb{R}^n$ and any $\gamma, M \geq 0$

$$\begin{aligned}
&\mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, [(b(\theta)+1)M_2 \hat{r}^2 + (b(\theta)+2)M\sigma^2]^{1/2}\big) \\
&\qquad \leq H_1 \big(\tfrac{en}{|I_o|}\big)^{-\alpha_1 |I_o|} e^{-m_1 M} + H_1' \big(\tfrac{en}{|I_*|}\big)^{-\alpha_1' |I_*|} e^{-m_1'' M} \leq H_2 e^{-m_2 M}, \\
&\mathbb{P}_\theta\big(\hat{r}^2 \geq \sigma^2 (M_0' + \gamma) |I_o| \log(\tfrac{en}{|I_o|}) + (M+1)\sigma^2\big) \leq H_0' (\tfrac{ne}{|I_o|})^{-\gamma |I_o|} e^{-M},
\end{aligned}$$

where all the other constants $(H_1, m_1, H_1', \alpha_1', M_0', H_0')$ are defined in Theorems 2.1 and 2.2. Notice that the above size relation holds uniformly in $\theta \in \mathbb{R}^n$. Although the coverage relation is also uniform in $\theta \in \mathbb{R}^n$, the main (unavoidable) problem is the dependence of the coverage relation on $b(\theta)$. That is why we introduced the EBR condition which essentially provides control over the quantity $b(\theta)$.

Remark 2.5. *The smaller constant τ_0 (involved in the definition of the EBR condition) is, the less restrictive the EBR condition is, the limiting case $\tau_0 \downarrow 0$ corresponds basically to no condition. However, the main message here is that for any specific distribution of error vector ξ there is always some value of the constant τ_0 in the EBR condition (bounded away from zero, depending on how "bad" ξ is). We treat a general situation and are not concerned with the most exact (smallest) value for τ_0, our bound for τ_0 is in terms of α and possibly too conservative for each specific distribution of ξ.*

The idea of the set $I_* = I_o^{\tau_0}$ introduced by (2.21) is that it contains $i_* = |I_*|$ most significant coordinates of vector θ that are (essentially) not missed by the procedure $\hat{I}$. The bias term of the rate $r^2(I_*, \theta)$ is the error that is made when setting significant coordinates to zero (whereas they may not be zero). Large ratio $b(\theta)$ defined by (2.26) means that this error is relatively large as compared to the variance part of the rate $r^2(I_*, \theta)$. In a way, such θ's "trick" the procedure $\hat{\theta}$ and can therefore be regarded as deceptive. For each $\theta \in \mathbb{R}^n$, $b(\theta)$ measures the amount of deceptiveness of θ: the bigger $b(\theta)$, the more deceptive θ. The EBR condition says that the deceptiveness has to be restricted: $\Theta_{\mathrm{eb}}(t) = \{\theta \in \mathbb{R}^n : b(\theta) \leq t\}$. An explicit example of EBR parameters is the set of *self-similar* parameters introduced in [86] which is in our terms $\Theta_{\mathrm{ss}}(p, c, \tau_0) = \{\theta \in \ell_0[p] : |I_*(\theta)| \geq cp\}$ for $p \in [n]$, $c \in (0, 1]$. If $\theta \in \Theta_{\mathrm{ss}}(p, c, \tau_0)$, then $p \leq c^{-1}|I_*|$ and $\sum_{i \in I_*^c} \theta_i^2 \leq \sum_{|I_*|}^{c^{-1}|I_*|} \theta_{[i]}^2 \leq (c^{-1} - 1)\tau_0 \sigma^2 |I_*| \log(\frac{en}{|I_*|})$, where the second inequality follows by the oracle definition. Hence, $\Theta_{\mathrm{ss}}(p, c, \tau_0) \subseteq \Theta_{\mathrm{eb}}((c^{-1} - 1)\tau_0)$. Notice that $\Theta_{\mathrm{ss}}(p, c, \tau) \subseteq \Theta_{\mathrm{eb}}((c^{-1} - 1)\tau)$ for any $\tau > 0$.

In particular, for $\theta \in \Theta_{\mathrm{ss}}(p, 1, \tau_0) = \{\theta \in \ell_0[p] : |I_*(\theta)| = p\}$, the insignificant coordinates I_*^c of such θ's are the true zeros and the significant coordinates I_* are sufficiently distinct from zero. Then the set $I_*(\theta)$ coincides with the support $\mathrm{supp}(\theta) \triangleq \{i \in [n] : \theta_i \neq$

$0\}$, i.e., $I_*(\theta) = \mathrm{supp}(\theta)$, so that $B(I_*,\theta) = 0$, implying $\Theta_{\mathrm{ss}}(p,1,\tau_0) \subseteq \Theta_{\mathrm{eb}}(0)$. This class consists of the "nicest" (least deceptive) parameters satisfying the EBR condition with zero deceptiveness $t = 0$. The uncertainty quantification result is the strongest for this class because the inflating factor is the smallest as $t = 0$. More about the EBR condition is in Section 2.3.

2.2.4. CONFIDENCE BALL OF $n^{1/4}$-RADIUS WITHOUT EBR

By analyzing the previous results, we see that the resulting radius $\hat{r}$ of our constructed confidence ball is of the oracle rate only under the EBR condition. In general, $\hat{r}^2$ underestimates the oracle rate $r^2(\theta)$. The difference is the bias term which may in general be of a bigger order than the variance part, leading to a bad coverage. Suppose we want to construct a confidence ball of a full coverage uniformly over the whole space $\mathbb{R}^n$. Recall however that, in view of the above mentioned negative results of [60], [28], [7] and [67], no data dependent ball can provide full coverage and adaptive size simultaneously. Insisting on the full coverage, one can at best adapt to sparsity levels only in the range $|I_*(\theta)| \geq C\sqrt{n}$ (i.e., actually for non-sparse parameters) and the term of order $\sigma^2\sqrt{n}$ should be in the radius. Let us give a heuristics behind this. An idea is to mimic the quantity $\|\theta - \hat{\theta}\|^2$ by $\hat{R}^2 = \|X - \hat{\theta}\|^2$. Clearly, there is a lot of bias in $\hat{R}^2$, the biggest part of which is due to the term $\sigma^2\|\xi\|^2$ contained in $\hat{R}$. To de-bias for that part, we need to subtract its expectation $\sigma^2\mathbb{E}_\theta\|\xi\|^2 = n\sigma^2$, where we assumed $\mathrm{Var}(\xi_i) = 1$ in the model (2.1) for simplicity. However, even de-biased quantity $\hat{R}^2$ can only be controlled up to a margin of the order $\sigma^2\sqrt{n}$. That is why a term of the order $\sigma n^{1/4}$ is necessary in the radius of the confidence ball to provide coverage uniformly over the whole space $\mathbb{R}^n$.

To handle some technical issues in this case, we impose the following additional condition.

CONDITION (B2). Besides X given by (2.1), we also observe $X' \in \mathbb{R}^n$ independent of X, where $X' = \theta + \sigma\xi'$, the random vector ξ' satisfies the following relations:

$$\begin{aligned} &\mathbb{P}_\theta\big(|\langle v,\xi'\rangle| \geq \sqrt{M}\big) \leq \psi_1(M) \ \ \forall\, v \in \mathbb{R}^n : \|v\| = 1; \\ &\mathbb{P}_\theta\big(\big|\|\xi'\|^2 - V(X')\big| \geq M\sqrt{n}\big) \leq \psi_2(M), \quad \text{for some statistic } V(X'). \end{aligned} \tag{B2}$$

Here $\psi_1(M), \psi_2(M)$ are some positive monotonically decreasing functions such that $\psi_1(M) \downarrow 0$ and $\psi_2(M) \downarrow 0$ as $M \uparrow \infty$.

Typically, $\mathbb{E}\xi'_i = 0$, $\mathrm{Var}(\xi'_i) = 1$, $i \in [n]$, then $V(X') = n$. Condition (B2) is satisfied for independent normals $\xi_i \overset{\mathrm{ind}}{\sim} \mathrm{N}(0,1)$ even if we do not have the sample X' at our disposal. Indeed, in this case we can "duplicate" the observations by randomization at the cost of doubling the variance in the following manner: create samples $X' = X + \sigma Z$ and $X'' = X - \sigma Z$, for a $Z = (Z_1,\dots,Z_n)$ (independent of X) such that $Z_i \overset{\mathrm{ind}}{\sim} \mathrm{N}(0,1)$. Relations (B2) are then fulfilled with exponential functions $\psi_l(M) = Ce^{-cM}$, $l = 1,2$, for some $C, c > 0$. If the sub-gaussianity condition (1.20) is fulfilled for ξ' (which is the same as Condition (B0) in case of independent ξ'_i's), then $\psi_1(M) = e^{-\rho M}$. By Chebyshev's inequality, we see that the second relation in (B2) is fulfilled with function $\psi_2(M) = cM^{-2}$ for any zero mean independent ξ'_i's with $\mathbb{E}_\theta\xi'^4_i \leq C$.

Coming back to the problem of constructing a confidence ball of full coverage uniformly over $\mathbb{R}^n$, let $\hat{\theta}$ and $\hat{I}$ be defined as before and based on the sample X. We propose

to mimic $\|\theta-\hat\theta\|^2$ by the de-biased quantity $\|X'-\hat\theta\|^2-\sigma^2V(X')$ plus additional $\sigma^2\sqrt{n}$-order term to control its oscillations, leading us to the following data dependent radius

$$\tilde R_M^2=\big(\|X'-\hat\theta\|^2-\sigma^2V(X')+2\sigma^2G_M\sqrt{n}\big)_+,\quad\text{where}\quad G_M=\sqrt{M(M+M_1)},\tag{2.27}$$

$x_+=x\vee0$ and the constant M_1 is from Theorem 2.1. The next theorem establishes the coverage and size properties of the confidence ball $B(\hat\theta,\tilde R_M)$.

Theorem 2.4. *Let Conditions* (B0), (B1) *and* (B2) *be fulfilled and* $\tilde R_M^2$ *be defined by* (2.27). *Then for any* $M\ge0$

$$\sup_{\theta\in\mathbb{R}^n}\mathbb{P}_\theta\big(\theta\notin B(\hat\theta,\tilde R_M)\big)\le\psi_1(M/4)+\psi_2(M)+H_1e^{-m_1M},$$
$$\sup_{\theta\in\mathbb{R}^n}\mathbb{P}_\theta\big(\tilde R_M^2\ge g_M(\theta,n)\big)\le\psi_1(M/4)+\psi_2(M)+2H_1e^{-m_1M},$$

$g_M(\theta,n)=M_1r^2(\theta)+M\sigma^2+4\sigma^2G_M\sqrt{n}$ *and the constants* H_1,m_1,M_1 *are defined in Theorem 2.1.*

The proof of this theorem is exactly in the same way as the proof of Theorem 5.4, with the only difference that everywhere (in the proof of Theorem 5.4) Y,Y' and N should be read as X,X' and n, also Condition (A4) and Theorem 5.1 should be read as Condition (B2) and Theorem 2.1.

By taking large enough M we can ensure the coverage and size relations uniformly over the entire space $\mathbb{R}^n$. However, notice the price for this overall uniformity: the radius of the constructed confidence ball is essentially of the order $\sigma n^{1/4}+r(\theta)$. So, it is always of the order at least $\sigma n^{1/4}$ even for very sparse parameters θ, and it is of the oracle rate order only for non-sparse parameters, when $r(\theta)\ge C\sigma n^{1/4}$. This is a fundamental problem for uncertainty quantification, which typically occurs when the parameter θ has many smallish coordinates θ_i, say, with θ_i^2 just under the noise level σ^2. Clearly, in this case no method can reliably assign those coordinates to the significant set. As demonstrated in [60], [28], [7] and [67], the above mentioned price in the form of a big radius is absolutely unavoidable (even in the case of just two sparsity classes as is shown in [67]), as soon as we require uniform coverage.

2.2.5. IMPLICATIONS: THE MINIMAX RESULTS OVER SPARSITY CLASSES

In this section we elucidate the potential strength of the local approach. In particular, we demonstrate how the global adaptive minimax results over certain scales can be derived from the local results. Note that the oracle rate $r(\theta)$ is a local quantity in that it quantifies the level of accuracy of inference about specific θ and originally it is not linked to any particular scale of classes. As we mentioned in Section 1.2, it is always possible to relate the oracle rate to various scales. Precisely, if we want to establish global adaptive minimax results over certain scale, say, $\{\Theta_\beta,\beta\in\mathcal{B}\}$, with corresponding minimax rates $\{r(\Theta_\beta),\beta\in\mathcal{B}\}$ (the minimax rate over Θ_β is $r^2(\Theta_\beta)\triangleq\inf_{\hat\theta}\sup_{\theta\in\Theta_\beta}\mathbb{E}_\theta\|\hat\theta-\theta\|^2$, where the infimum is taken over all estimators), the only thing we need to show is

$$\sup_{\theta\in\Theta_\beta}r^2(\theta)\le cr^2(\Theta_\beta),\quad\text{for all }\beta\in\mathcal{B}.$$

If the above property holds, we say the oracle rate $r(\theta)$ *covers* the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$. In this case, the local results on the estimation, the posterior contraction and the size relation of the confidence ball will immediately imply the corresponding global adaptive minimax results over the covered scale, (actually, simultaneously for all scales that are covered by the oracle rate $r(\theta)$). As to the coverage property, according to Theorem 2.3, it holds uniformly only over the EBR class $\Theta_{\mathrm{eb}}(t)$, whichever scale we consider. Thus, specializing the coverage property to a particular scale boils down to intersecting this scale with the EBR class $\Theta_{\mathrm{eb}}(t)$ in the coverage property.

Next, we consider two sparsity scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ for which the adaptive minimax results (on the estimation problem, the contraction rate of the empirical Bayes posterior, and the size property of the confidence ball $B(\hat{\theta}, (M_2\hat{r}^2 + M)^{1/2})$) follow from our local results Theorems 2.1 and 2.3. The results for other (covered) scales can also be readily derived.

Nearly black vectors. For $p_n \in [n]$ such that $p_n = o(n)$ as $n \to \infty$ (we use the usual o, O notation to describe the asymptotic behavior of certain quantities as $n \to \infty$), introduce the sparsity class $\ell_0[p_n] = \{\theta \in \mathbb{R}^n : s(\theta) = |I^*(\theta)| \le p_n\}$, where by $I^*(\theta)$ and $s(\theta)$ we denote the active index set and the sparsity of $\theta \in \mathbb{R}^n$:

$$I^*(\theta) = \{i \in [n] : \theta_i \neq 0\}, \qquad s(\theta) = |I^*(\theta)|. \tag{2.28}$$

The minimax estimation rate over the class of nearly black vectors $\ell_0[p_n]$ with the sparsity parameter p_n is known to be $r^2(\ell_0[p_n]) = O\big(\sigma^2 p_n \log(\frac{n}{p_n})\big)$ as $n \to \infty$; see [36]. By the definition (2.17) of the oracle rate $r^2(\theta)$, we have that $r^2(\theta) \le r^2(I^*(\theta), \theta)$. Then we obtain trivially that

$$\sup_{\theta \in \ell_0[p_n]} r^2(\theta) \le \sup_{\theta \in \ell_0[p_n]} r^2(I^*(\theta), \theta) \le \sigma^2 p_n \log\big(\tfrac{en}{p_n}\big) = O\big(r^2(\ell_0[p_n])\big).$$

The last relation, Theorems 2.1 and 2.3 immediately imply the adaptive minimax results for the scale $\ell_0[p_n]$. We summarize these results in the following corollary.

Corollary 2.1. *Under the conditions of Theorem 2.3, we have for any $M \ge 0$*

$$\sup_{\theta \in \ell_0[p_n]} \mathbb{E}_\theta \hat{\pi}\big(\|\vartheta - \theta\|^2 \ge M_0 \sigma^2 p_n \log(\tfrac{en}{p_n}) + M\sigma^2 | X\big) \le H_0 e^{-m_0 M},$$
$$\sup_{\theta \in \ell_0[p_n]} \mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 \ge M_1 \sigma^2 p_n \log(\tfrac{en}{p_n}) + M\sigma^2\big) \le H_1 e^{-m_1 M},$$
$$\sup_{\theta \in \ell_0[p_n]} \mathbb{P}_\theta\big(\hat{r}^2 \ge M_0' \sigma^2 p_n \log(\tfrac{en}{p_n}) + (M+1)\sigma^2\big) \le H_0' e^{-M}.$$

The next theorem describes some "over-dimensionality" (or "undersmoothing") control of the empirical Bayes posterior $\hat{\pi}(I|X)$ from the $\mathbb{P}_\theta$-perspective.

Theorem 2.5. *Let $s(\theta)$ be defined by (2.28). Under the conditions of Theorem 2.2, there exist $M_4, m_4 > 0$ such that for any $M > M_4$ and $\theta \in \mathbb{R}^n$*

$$\mathbb{E}_\theta \hat{\pi}(I : |I| > M s(\theta) | X) \le C_0 \exp\big\{-m_4 s(\theta)\big[(M - M_4)\log(\tfrac{en}{s(\theta)}) - M \log M\big]\big\}.$$

In particular, there exist constants $M_4', m_4' > 0$ such that

$$\mathbb{E}_\theta \hat{\pi}(I : |I| > M_4' s(\theta) | X) \le C_0 \exp\{-m_4' s(\theta) \log(\tfrac{en}{s(\theta)})\}.$$

The above theorem is a local type result, but can readily be specialized to the sparsity class $\theta \in \ell_0[p_n]$ in the minimax sense. If $s(\theta) \ge 1$, the probability bound goes to 0 as $n \to \infty$.

Weak ℓ_q-balls. For $q \in (0,2)$, the *weak ℓ_q-ball* with the sparsity parameter p_n is defined by

$$m_q[p_n] = \{\theta \in \mathbb{R}^n : \theta_{[i]}^2 \le (p_n/n)^2 (n/i)^{2/q}, i \in \mathbb{N}_n\}, \quad p_n = o(\sigma n) \text{ as } n \to \infty,$$

where $\theta_{[1]}^2 \ge \ldots \ge \theta_{[n]}^2$ are the ordered $\theta_1^2, \ldots, \theta_n^2$. This scale can be thought of as Sobolev hyper-rectangle for ordered (with unknown locations) coordinates: $m_q[p_n] = \mathcal{H}(\beta, \delta_n) = \{\theta \in \mathbb{R}^n : |\theta_{[i]}| \le \delta_n i^{-\beta}\}$, with $\delta_n = n^{1/q} \frac{p_n}{n}$ and $\beta = 1/q > 1/2$.

Denote $j = O_\theta(i)$ if $\theta_i^2 = \theta_{[j]}^2$, with the convention that in the case $\theta_{i_1}^2 = \ldots = \theta_{i_k}^2$ for $i_1 < \ldots < i_k$ we let $O_\theta(i_{l+1}) = O_\theta(i_l) + 1$, $l = 1, \ldots, k-1$. The minimax estimation rate over this class is $r^2(m_q[p_n]) = n(\frac{p_n}{n})^q [\sigma^2 \log(\frac{n\sigma}{p_n})]^{1-q/2}$ when $n^{2/q}(\frac{p_n}{n})^2 \ge \sigma^2 \log n$, and $r^2(m_q[p_n]) = n^{2/q}(\frac{p_n}{n})^2 + \sigma^2$ when $n^{2/q}(\frac{p_n}{n})^2 < \sigma^2 \log n$, as $n \to \infty$; see [38] and [23]. Then take $I^*(\theta) = \{i \in \mathbb{N}_n : O_\theta(i) \le p_n^*\}$, with $p_n^* = en(\frac{p_n}{n\sigma})^q [\log(\frac{n\sigma}{p_n})]^{-q/2}$ in the case $n^{2/q}(\frac{p_n}{n})^2 \ge \sigma^2 \log n$, to derive

$$\begin{aligned}
\sup_{\theta \in m_q[p_n]} r^2(\theta) &\le \sup_{\theta \in m_q[p_n]} r^2(I^*(\theta), \theta) \le \sigma^2 p_n^* \log(\tfrac{en}{p_n^*}) + n^{2/q}(\tfrac{p_n}{n})^2 \sum_{i > p_n^*} i^{-2/q} \\
&\le K_1 \sigma^2 p_n^* \log(\tfrac{n\sigma}{p_n}) + K_2 n^{2/q} (\tfrac{p_n}{n})^2 (p_n^*)^{1-2/q} \\
&\le K n (\tfrac{p_n}{n})^q \big[\sigma^2 \log(\tfrac{n\sigma}{p_n})\big]^{1-q/2} = O\big(r^2(m_q[p_n])\big), \qquad (2.29)
\end{aligned}$$

for some $K = K(q)$. The case $n^{2/q}(\frac{p_n}{n})^2 < \sigma^2 \log n$ is treated similarly by taking $p_n^* = 0$.

Theorems 2.1 and 2.3 imply the minimax adaptive results for the scale $m_q[p_n]$.

Corollary 2.2. *Under the conditions of Theorem 2.3, we have for any $M \ge 0$*

$$\begin{aligned}
&\sup_{\theta \in m_q[p_n]} \mathbb{E}_\theta \hat{\pi}\big(\|\vartheta - \theta\|^2 \ge M_0 K r^2(m_q[p_n]) + M\sigma^2 | X\big) \le H_0 e^{-m_0 M}, \\
&\sup_{\theta \in m_q[p_n]} \mathbb{P}_\theta \|\big(\hat{\theta} - \theta\|^2 \ge M_1 K r^2(m_q[p_n]) + M\sigma^2\big) \le H_1 e^{-m_1 M}, \\
&\sup_{\theta \in m_q[p_n]} \mathbb{P}_\theta\big(\hat{r}^2 \ge M_0' K r^2(m_q[p_n]) + (M+1)\sigma^2\big) \le H_0' e^{-M}.
\end{aligned}$$

The following theorem concerns the "over-dimensionality" control for the class $m_q[p_n]$.

Theorem 2.6. *Let the conditions of Theorem 2.2 be fulfilled, $p_n^* = en(\frac{p_n}{n\sigma})^q \big[\log(\frac{n\sigma}{p_n})\big]^{-q/2}$ and $n^{2/q}(\frac{p_n}{n})^2 \ge \sigma^2 \log n$. Then there exist constants $M_5, m_5 > 0$ such that for any $M > M_5$ there exists $n_0 = n_0(M, q)$ such that, for all $n \ge n_0$,*

$$\sup_{\theta \in m_q[p_n]} \mathbb{E}_\theta \hat{\pi}(I : |I| > M p_n^* | X) \le C_0 \exp\{-m_5 (M - M_5) p_n^* \log(\tfrac{n\sigma}{p_n})\}.$$

Notice that the exponential upper bound from the last relation converges to zero as $n \to \infty$ because $p_n^* \log(\frac{n\sigma}{p_n}) \geq e(\sigma^2 \log n)^{q/2} (\log(\frac{n\sigma}{p_n})^{1-q/2}$.

Remark 2.6. *The same minimax results hold over the so called* strong ℓ_q-ball $\ell_q[p_n] = \{\theta \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^n |\theta_i|^q \leq (\frac{p_n}{n})^q\}$, $q \in (0,2)$, *since* $\ell_q[p_n] \subseteq m_q[p_n] \subseteq \ell_{q'}[p_n]$ *for any* $q' > q$.

2.3. THE EBR CONDITION

A new perspective on EBR – the EBR scale. As we mentioned in the introduction of this thesis, it is impossible to construct optimal (fully) adaptive confidence set in the minimax sense over traditional smoothness and sparsity scales with a prescribed high coverage probability. Namely, there exist "deceptive" parameters $\theta \in \Theta_0' = \mathbb{R}^n \backslash \Theta_0$ for which the coverage property in (1.14) may not hold for arbitrarily small α_1. Removing deceptive parameters Θ_0' and restricting to the remaining set Θ_0 of non-deceptive parameters resolves this issue. This was the original motivation of introducing the EBR condition.

An interesting advantageous feature of the EBR condition introduced in this chapter is that it leads to *slicing of the entire parameter space* $\mathbb{R}^n$. This opens up a new perspective on the EBR and its role in the deceptiveness issue, which we explain next.

Note that the EBR condition $\theta \in \Theta_{\text{eb}}(t,\tau)$ (see (2.25)) is actually a family of embedded conditions parametrized by $t \geq 0$: $\Theta_{\text{eb}}(t_1,\tau) \subseteq \Theta_{\text{eb}}(t_2,\tau)$ for $t_1 \leq t_2$ and any $\tau \geq 0$. Note that, by the oracle definition, $\Theta_{\text{eb}}(\tau n,\tau) = \mathbb{R}^n$ for any $\tau > 0$. An important observation is that this family of conditions effectively introduces a new scale $\cup_{t\geq 0}\Theta_{\text{eb}}(t,\tau) = \cup_{0\leq t\leq \tau n}\Theta_{\text{eb}}(t,\tau)$ (for any fixed $\tau > 0$), to be called the *EBR scale*, with the structural parameter $t \geq 0$ measuring the allowed amount of deceptiveness for parameters $\theta \in \Theta_{\text{eb}}(t,\tau)$. Indeed, this scale "slices" $\mathbb{R}^n$ in the sense that $\mathbb{R}^n = \cup_{0\leq t\leq \tau n}\Theta_{\text{eb}}(t,\tau)$. The main benefit of introducing the EBR scale is that it gives the slicing of the entire space that is very suitable for uncertainty quantification. Indeed, the dictum "removing deceptive parameters" becomes a very natural notion in terms of the scale $\cup_{0\leq t\leq \tau n}\Theta_{\text{eb}}(t,\tau)$ as it is nothing else but restricting the amount of deceptiveness t. This provides a new perspective at the above mentioned "deceptiveness" issue: basically, each parameter $\theta \in \mathbb{R}^n$ has a certain amount of "deceptiveness" that is measured by the excessive bias ratio $b_\tau(\theta)$ defined by (2.26), or the smallest t for which $\theta \in \Theta_{\text{eb}}(t,\tau)$. The larger t, the more deceptive parameters are allowed in $\Theta_{\text{eb}}(t,\tau)$. A mild and controllable price for the uniformity over $\Theta_{\text{eb}}(t,\tau)$ in the coverage relation is the amount of inflating of the confidence ball needed to provide a guaranteed high coverage for the parameters of deceptiveness at most t. We should mention that the EBR scale is intrinsically tied to our Bayesian procedure as it depends on the proposed family of priors $\{\pi_I, I \in \mathcal{I}\}$. A different family may lead to a different EBR scale. Note however that a version of our EBR condition (adapted to the sparsity scale) is still used for a different (horseshoe) prior in paper [86] .

Interestingly, slicing is also possible by the parameter $\tau > 0$: $\mathbb{R}^n = \cup_{\tau\geq 0}\Theta_{\text{eb}}(t,\tau)$ (for any $t > 0$), the embedding goes in the opposite direction: the smaller the τ, the weaker the EBR. Namely, $\Theta_{\text{eb}}(t,\tau_2) \subseteq \Theta_{\text{eb}}(t,\tau_1)$ for any $0 \leq \tau_1 \leq \tau_2$, $t > 0$, and the "limiting" EBR set $\lim_{\tau\downarrow 0}\Theta_{\text{eb}}(t,\tau)$ expands to the entire space: $\Theta_{\text{eb}}(t,0) = \mathbb{R}^n$. Besides, notice that the inflating factor in the confidence ball from Theorem 2.3 will not not increase as $\tau \downarrow 0$ (in fact, it will decrease). A paradox seems to have emerged: by considering very small τ's,

we can have less deceptiveness without any price in the coverage relation. However, this paradox is resolved by reminding that the coverage relation from Theorem 2.3 does not hold for arbitrarily small τ_0 because in (2.21) $\tau_0 > \frac{1+\rho}{1-\rho}\bar{\tau}$, showing that "there is no free lunch". The lower bound for τ_0 can be relaxed (made smaller) for specific distribution of ξ, but it will always be some positive threshold reflecting the complexity of ξ.

The EBR does not affect the minimaxity over the sparsity scale $\ell_0[p]$. The EBR condition is mild from the minimax point of view in the following sense: if we take the traditional sparsity class $\ell_0[p] = \{\theta \in \mathbb{R}^n : |I^*(\theta)| = \|\theta\|_0 \le p\}$ for $p \in \mathbb{N}_n$ and remove non-EBR parameters, then the minimax rate over the remaining part will not change (up to a constant). We outline the argument below. The minimax estimation rate was established in [23] (Theorem 4 from [23], formulated in our notation): for some universal constant $c > 0$,

$$r^2(\ell_0[p]) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \ell_0[p]} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \ge c\sigma^2 p \log(en/p).$$

The proof is based on considering the subset $\mathcal{B}_1(p) = \{\theta \in \mathbb{R}^n : |I^*(\theta)| \le p, |\theta_i| \le \sigma^2 \log(en/p)\} \subset \ell_0[p]$ and establishing the required lower bound for the minimax risk $R(\mathcal{B}_1(p))$ over the set $\mathcal{B}_1(p)$, thus obtaining $r^2(\ell_0[p]) \ge r^2(\mathcal{B}_1(p)) \ge c\sigma^2 p \log(en/p)$. Inspecting all the steps in the proof, we see that essentially the same lower bound (with a different constant c) holds for another subset of $\ell_0[p]$: $\mathcal{B}_2(p,\tau_0) = \{\theta \in \mathbb{R}^n : |I^*(\theta)| = p, 2\tau_0\sigma^2 \log(en/p) \le |\theta_i| \le (2\tau_0+1)\sigma^2 \log(en/p)$ for all $i \in I^*(\theta)\}$, for τ_0 defined in (2.21). For each $\theta \in \mathcal{B}_2(p,\tau_0)$, we have $I_*(\theta) = I_o^{\tau_0}(\theta) = I^*(\theta)$ so that $I_*^c(\theta) = \varnothing$, $|I_*(\theta)| = p$, and the EBR condition is trivially satisfied for any $t \ge 0$:

$$\frac{\sum_{i \in I_*^c} \theta_i^2}{\sigma^2(1 + |I_*| \log(en/|I_*|))} = 0 \le t.$$

This means that $r^2(\ell_0[p] \cap \Theta_{\mathrm{eb}}(t)) \ge r^2(\mathcal{B}_2(p,\tau_0)) \ge c\sigma^2 p \log(en/p)$.

2.4. SIMULATIONS

Here we present a small simulation study. In the model (2.1), we used $n = 500$, $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, $\sigma = 1$, and signals $\theta = (\theta_1, \dots, \theta_n)$ of the form $\theta = \theta(p,A) = (0, \dots, 0, A, \dots, A) = (0 \cdot 1_{n-p}, A1_p)$, where $1_p = (1, \dots, 1) \in \mathbb{R}^p$. The first $n-p$ zero coordinates are "insignificant" and the last p coordinates of value A are "significant". Different sparsity levels $p \in \{25, 50, 100\}$ and "signal strengths" $A \in \{3, 4, 5\}$ are considered. It is easy to compute the oracle $I_o(\theta(p,A)) = \{n-p+1, \dots, n\}$ and the oracle rate $r^2(\theta(p,A)) = B(I_o,\theta) + V(I_o) = V(I_o) = |I_o| \log(en/|I_o|) = p \log(en/p)$. Then the excessive bias $B(I_o, \theta(p,A)) = 0$, so that the deceptiveness $b(\theta(p,A),1) = 0$ and hence the EBR condition (in terms of the standard oracle I_o) is satisfied with $t = 0$: $\theta(p,A) \in \Theta_{\mathrm{eb}}(0,1)$ for all considered $\theta(p,A)$.

In case $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, Condition (B0) is fulfilled with $\alpha = 0.4$, leading to $\kappa > 3.24$ in Condition (B1). The bound for κ coming from Condition (B1) is typically too conservative as it is for the general situation of unknown distribution of ξ. For example, if $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, Condition (B1) can be relaxed to $\kappa \ge 2.04$; see Section 2.5. It is desirable to choose κ in a

data dependent way. In our simulation study, we choose κ via a *cross-validation* procedure. For that, we create two independent normal samples $X'_i = X_i + \eta_i$ and $X''_i = X_i - \eta_i$, where simulated independent standard normal η_i's are assumed to be independent of ξ_i, $i \in [n]$. Then X'_i and X''_i are independent random variables with means θ_i and variances 2. In words, the observation sample can be duplicated at the cost of multiplying the variance by 2. Using these samples, we estimate $\kappa > 0$ as follows: let $\check{\theta}' = \check{\theta}'(\hat{I}') = (X'_i 1\{i \in \hat{I}'\}, i \in [n])$, where $\hat{I}' = \hat{I}'(\kappa) = \operatorname{argmin}_{I \in \mathcal{I}} \Big\{ -\sum_{i \in I} (X'_i)^2 + 2(2\kappa+1)|I| \log\big(\frac{en}{|I|}\big) \Big\}$, then $\hat{\kappa} = \operatorname{argmin}_{\kappa \in (0, \log n]} \|\check{\theta}'(\hat{I}'(\kappa)) - X''\|^2$. Let $\hat{I} = \hat{I}(\hat{\kappa})$ be defined by (2.12) with $\kappa = \hat{\kappa}$, and $\check{\theta} = X(\hat{I})$ be defined by (2.13).

Now consider the confidence ball $B\big(\check{\theta}, \hat{R}\big)$, where $\hat{R} = \bar{M}[(\hat{b}+1)_+ \hat{r}^2]^{1/2}$, $\hat{r}^2 = |\hat{I}| \log(en/|\hat{I}|)$ given by (2.24), and $\hat{b} = \frac{\sum_{i \in \hat{I}^c} (X_i^2 - 1)}{|\hat{I}| \log(en/|\hat{I}|)}$ is an estimate of the deceptiveness $b(\theta, 1)$ defined by (2.26). The construction of the radius $\hat{R}$ is inspired by the local result formulation in Remark 2.4. The quantity $[(\hat{b}+1)_+ \hat{r}^2]^{1/2}$ is an empirical counterpart for the oracle rate $r(\theta)$, but even the oracle rate radius needs to be inflated to ensure coverage. The multiplicative factor $\bar{M}$ is intended to trade-off the size of the ball against its coverage probability. Theoretical inflating factor from Theorem 2.3 is too conservative, as it is for the general situation of Condition (B0). In this simulation study it is enough to take $\bar{M} = \sqrt{2}$, thus $\hat{R} = [2(\hat{b}+1)_+ \hat{r}^2]^{1/2}$, which yielded good results for all the cases.

Table 2.1 shows the performance of the confidence ball $B\big(\check{\theta}, \hat{R}\big)$. For each $\theta = \theta(p, A)$, with $p \in \{25, 50, 100\}$ and $A \in \{3, 4, 5\}$, we simulated 100 data vectors X from the model (2.1) and computed two quantities: 1) the ratio $\overline{\hat{R}}/r(\theta)$ of the average of the radius $\overline{\hat{R}}$ to the oracle rate $r(\theta)$ defined by (2.17); 2) the frequency $\bar{\alpha}$ of the event that confidence ball $B(\check{\theta}, \hat{R})$ contains the signal θ. The former characterizes the size of the confidence ball $B(\check{\theta}, \hat{R})$ relative to the oracle rate, and the latter its coverage. What one can conclude from the table: the higher the signal strength, the smaller the ratio $\overline{\hat{R}}/r(\theta)$; uncertainty quantification does not seem to benefit from the sparsity in terms of relative size (with respect to the oracle rate) and coverage. Of course, the absolute size is certainly better for more sparse signals as the oracle rate is then smaller. In particular, $r(\theta(25, A)) = 9.99$, $r(\theta(50, A)) = 12.85$ and $r(\theta(100, A)) = 16.15$, $A = 3, 4, 5$.

Table 2.1: The ratio $\overline{\hat{R}}/r(\theta)$ and the frequency $\bar{\alpha}$ of the event that the confidence ball $B(\check{\theta}, \hat{R})$ contains the signal $\theta(p, A)$ computed for 100 vectors X simulated from (2.1) with $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, $n = 500$ and $\sigma = 1$.

p	25			50			100		
A	3	4	5	3	4	5	3	4	5
$\overline{\hat{R}}/r(\theta)$	1.81	1.71	1.51	1.78	1.5	1.4	1.51	1.34	1.34
$\bar{\alpha}$	0.98	0.97	0.95	0.99	0.97	1	0.97	1	1

At the first site surprisingly, the constructed confidence ball $B(\check{\theta}, \hat{R})$ appeared to perform well also for deceptive parameters, like $\bar{\theta} = \bar{\theta}(\delta, p, A) = (\delta 1_{n-p}, A 1_p)$, with p, A as before and $\delta > 0$. We get very similar (good) results as in Table 2.1 for *all* $\delta > 0$. Notice that the deceptivenesses may not be zero, for example, $b(\bar{\theta}(0.5, 25, A)) = 1.18$, $b(\bar{\theta}(0.5, 50, A)) = 0.68$, $b(\bar{\theta}(0.5, 100, A)) = 0.38$, $b(\bar{\theta}(0.8, 25, A)) = 3$, etc. The reason for

good results even for deceptive signals is that their oracle rate are large relative to $n^{1/4}$. Indeed, even the smallest oracle rate $r(\theta(0,25,3)) = 9.99 > (500)^{1/4} = 4.73$, which means that we are essentially in the $n^{1/4}$-situation of Theorem 2.4 rather than Theorem 2.3 when both optimal size and coverage are possible to attain. Basically, the signals $\bar{\theta}(\delta, p, A)$ are not sparse enough and/or the problem is not high-dimensional enough.

To see clearly the deceptiveness effect, we created a signal $\theta' = (0,\dots,0,A_1,\dots,A_p)$ of dimension $n = 500$, with sparsity $p = 10$, $A_i \overset{\text{ind}}{\sim} \mathrm{U}[0,4]$, $i = 1,\dots,p$. The oracle rate was $r(\theta') = 4.69 < (500)^{1/4} = 4.73$, the deceptiveness was $b(\theta') = 0.64$. Thus, this was a deceptive signal, but not in the $n^{1/4}$-situation anymore. The size was still very good $\overline{\hat{R}}/r(\theta) = 1.17$, but the coverage was pretty low $\bar{\alpha} = 0.53$. The deceptiveness manifested itself more prominently in the case $n = 5000$, $p = 10$, $A_i \overset{\text{ind}}{\sim} \mathrm{U}[0,4]$, $i = 1,\dots,p$. The oracle rate was $r(\theta') = 5.27 < (5000)^{1/4} = 8.41$, deceptiveness $b(\theta') = 1.7$. The size was still good $\overline{\hat{R}}/r(\theta) = 1.37$ as before, but the coverage was $\bar{\alpha} = 0.51$.

2.5. CONCLUDING REMARKS

Characterization of the significant and insignificant coordinates. The definition (2.17) of the oracle I_o implies the following characterization of the significant coordinates $\{\theta_i, i \in I_o\} = \{\theta_{[i]}, i = 1,\dots,i_o\}$:

$$\begin{aligned}
\theta^2_{[i_o]} &\ge \sigma^2\big[\log(\tfrac{en}{i_o}) - (i_o-1)\log(\tfrac{i_o}{i_o-1})\big],\\
\theta^2_{[i_o]} + \theta^2_{[i_o-1]} &\ge \sigma^2\big[2\log(\tfrac{en}{i_o}) - (i_o-2)\log(\tfrac{i_o}{i_o-2})\big],\\
&\dots,\\
\textstyle\sum_{i=1}^{i_o}\theta^2_{[i]} &\ge \sigma^2 i_o \log(\tfrac{en}{i_o}).
\end{aligned}$$

The insignificant coordinates $\{\theta_i, i \in I_o^c\} = \{\theta_{(i)}, i = 1,\dots,n-i_o\}$ can be characterized in a similar manner. Namely, from the definition of the oracle it follows that the insignificant coordinates $\{\theta_{(i)}, i = 1,\dots,n-i_o\}$ satisfy

$$\begin{aligned}
\theta^2_{(n-i_o)} &\le \sigma^2\big[\log(\tfrac{en}{i_o+1}) - i_o\log(\tfrac{i_o+1}{i_o})\big],\\
\theta^2_{(n-i_o)} + \theta^2_{(n-i_o-1)} &\le \sigma^2\big[2\log(\tfrac{en}{i_o+2}) - i_o\log(\tfrac{i_o+2}{i_o})\big],\\
&\dots,\\
\sum_{i=1}^{n-i_o}\theta^2_{(i)} &\le \sigma^2\big[n - i_o - i_o\log(\tfrac{n}{i_o})\big].
\end{aligned}$$

Improving constants. Since our approach applies to a very general situation, many constants involved in the conditions and proofs may be rather conservative. Indeed, we do not specify any distribution of ξ and even dot not assume independence of its coordinates. For the problem to be at all solvable, the vector ξ has to have some minimal structure which is in our case provided by Condition (B0). The constant $\alpha > 0$ reflects in a generic way how bad (or how good) the vector ξ is, implying that almost all the constants in the proofs and conditions depend on α. Clearly, if a distribution of ξ is specified, many bounds can be made more precise and many constants can be improved, including the

constants $\bar{\kappa}$ from Condition (B1) and τ_0 defined in (2.21), see Remark 2.5 for more detail on constant τ_0. Besides, some constants can be improved by using more precise inequalities at some steps of the proof. But this would make the presentation significantly lengthier without adding anything new conceptually.

For example, in case $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, we can sharpen up many constants in the proofs and conditions. In the proof of Lemma 2.1, we can compute exactly the right hand side of (2.30) by using the elementary identity: for $Y \sim \mathrm{N}(\mu_y, \sigma_y^2)$,

$$\mathbb{E}\exp\{\tfrac{aY^2}{2}\} = \exp\{\tfrac{a\mu_y^2}{2(1-a\sigma_y^2)} - \tfrac{1}{2}\log(1-a\sigma_y^2)\}, \quad \text{for any } a < \sigma_y^{-2}. \tag{S1}$$

By some tedious but straightforward calculations, we obtain the claim of Lemma 2.1 for any $h \in [0,1)$ with the constants $A_h = \frac{h}{2(1+h)}$, $B_h = \frac{h}{2(1-h)}$, $C_h = \frac{h}{2}$ and $D_h = \frac{h}{2} + \frac{1}{2}\log(1-h)$. If $I \setminus I_0 = \varnothing$, the bound holds also for $h = 1$ with $A_1 = \frac{1}{4}$, $B_1 = 0$, $C_1 = D_1 = \frac{1}{2}$. Next, since Lemma 2.1 now holds for any $h \in [0,1)$, we can try to optimize the choice of h in Lemma 2.2. We can also relax the requirement $c_1 > 2$ to $c_1 > 1$ in Lemma 2.2, leading to the bound for $\kappa \geq \bar{\kappa} = 2.04$.

The constants in the proof of Theorem 2.2 can also be improved in the normal case and we can use the bound $\mathbb{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2 = |I|^2 + 2|I| \leq 3|I|^2$ instead of (2.15) in the proof of Theorem 2.1.

Product prior. If, instead of the prior π, we take a prior $\bar{\pi} = \bar{\pi}_{K,\kappa} = \sum_{I\in\mathcal{I}} \lambda_I \pi_I$ with $\tau_i^2(I) = K\sigma^2 1\{i \in I\}$ for any fixed $K > 0$ (we can even allow $K = K_n \to \infty$, but $K_n = O(n)$, as $n \to \infty$) in (2.2) and $\lambda_I = c_{\kappa,n}\exp\{-\kappa|I|\log n\}$ (with $\kappa > \kappa_0$ for some $\kappa_0 > 0$) in (2.4), then all the results will hold with $\log n$ instead of $\log(\frac{en}{|I|})$ in the oracle rate (2.17). This case was studied in the first version of the arXiv-preprint of [13]. Thus, the results for the prior $\bar{\pi}$ are weaker than the results obtained in this chapter. For example, the minimax rates for the sparsity classes (Corollaries 2.1, 2.2) follow from these weaker results only if the sparsity parameter $p_n = O(n^\gamma)$ for $\gamma \in [0,1)$ as $n \to \infty$, otherwise we obtain only the *near-minimax* rates, with the factor $\log n$ instead of $\log(\frac{n}{p_n})$.

However, there is an advantageous feature of the prior $\bar{\pi}$ as compared with π. Namely, it is of the product structure: for $\lambda_I = c_\lambda \prod_{i\in I}\lambda_i$ with $c_\lambda = \prod_{i=1}^n (1+\lambda_i)^{-1}$, we compute $\bar{\pi} = \sum_{I\in\mathcal{I}}\lambda_I\pi_I = \bigotimes_{i=1}^n \big[\omega_i \mathrm{N}(\mu_{1,i}, K\sigma^2) + (1-\omega_i)\delta_0\big]$, $\omega_i = \frac{\lambda_i}{1+\lambda_i}$ ($\omega_i = \lambda(i \in I)$ is the prior probability that the random set I contains i). This leads to the product structure of the empirical Bayes posterior, so that the computation of the corresponding empirical Bayes estimator can easily be done in the coordinatewise fashion. Indeed, in our case $\lambda_i = \lambda = n^{-\kappa}$ and some computations give the following empirical Bayes posterior

$$\bar{\pi}(\vartheta|X) = \bigotimes_{i=1}^n \big[p_i \mathrm{N}\big(X_i, \tfrac{K\sigma^2}{K+1}\big) + (1-p_i)\delta_0\big], \qquad p_i = 1/\big[1 + h\exp\{-\tfrac{X_i^2}{2\sigma^2}\}\big],$$

where $p_i = \bar{\pi}(\theta_i \neq 0|X)$ and $h = h_{\kappa,K} = \frac{\sqrt{K+1}}{\lambda} = n^\kappa (K+1)^{1/2}$. The mean with respect to $\bar{\pi}(\vartheta|X)$ is readily obtained: $\bar{\theta} = \mathbb{E}_{\bar{\pi}}(\vartheta|X) = (p_i X_i,\, i \in [n])$, a shrinkage estimator with easily computable shrinkage factors p_i. Coordinatewise empirical Bayes medians can also be easily computed.

Cardinality dependent prior λ. Notice that the prior $\lambda = (\lambda_I, I \in \mathcal{I})$ defined by (2.4) depends on the set $I \in \mathcal{I}$ only via its cardinality $|I|$, i.e., $\lambda_I = g(|I|)$ for some nonnegative function $g(k)$, $k = 0, 1, \dots, n$. It is easy to see that in this case $\pi_n(k) = g(k)\binom{n}{k}$, $k = 0, 1, \dots, n$, determines the prior on the cardinality of I. Hence, the prior λ_I can always be modeled in two steps: first draw the random cardinality K according to the prior $\pi_n(k)$, and then given $K = k$, draw a random set I uniformly from the family of all subsets of $\mathcal{I}$ of cardinality k. Such priors λ are used in [33], where the cardinality prior $\pi_n(k)$ can be taken to be a so called "complexity prior" $\pi_n(k) = \exp\{-ak\log(bn/k)\}$ for some $a, b > 0$. Since $e^{k\log(n/k)} \le \binom{n}{k} \le e^{k\log(ne/k)}$, the resulting prior mass λ_I on I is bounded below and above by expressions of the type $\exp\{-a_1|I|\log(b_1 n/|I|)\}$, resembling the prior (2.4). The condition on the complexity prior from [33] essentially corresponds to our condition $\kappa > \bar{\kappa}$ for some $\bar{\kappa} > 0$ (Condition (B1)).

Connection of empirical Bayes with penalization method. Notice that the estimator $\check{\theta} = X(\hat{I})$ defined by (2.13) is the penalized estimator as introduced in [23] (cf. also [2]), where the selector ("estimator of the oracle") $\hat{I}$ is determined by the penalization criterion (2.12) with the penalty $P(I) = (2\kappa+1)\sigma^2|I|\log\left(\frac{en}{|I|}\right)$. This penalty is from the family of penalties corresponding to the *complete variable selection* case in [23] with the penalty constant $2\kappa+1$. Recall our rather specific choice of parameter $K_n(I)$ in (2.2) resulting in this penalty. As we mentioned, other choices of $K_n(I)$ are also possible, which would lead to other penalties. But the main term $\log\left(\frac{en}{|I|}\right)$ would always be present in the penalty because of the choice of prior λ_I. This reiterates the conclusion in [2] that essentially only this kind of penalties lead to adaptive penalized estimators.

Interestingly, recall that we require $\kappa > \bar{\kappa}$ for some $\bar{\kappa}$ (see Condition (B1)). For each specific situation, one can try to relax this bound to $\kappa > \kappa_0$, for some positive $\kappa_0 < \bar{\kappa}$, but κ_0 cannot be arbitrarily close to zero (in order for prior λ to sufficiently penalize big values of cardinality $|I|$). In a way, the condition $\kappa > \bar{\kappa} > 0$ (Condition (B1)) corresponds to the requirement in [23] (argued in [23] from a different perspective) that the penalty constant for penalized estimators should be bounded away from 1. It is argued in [23] that large penalty constants κ should also be avoided. We get the same conclusion by observing that the constants in claims (i)-(ii) of Theorem 2.1 become worse as $\kappa \to \infty$.

Computing the estimators. Note that the estimator (2.11) is a shrinkage estimator, and the estimator (2.13) is a hard thresholding procedure. Indeed, the estimator (2.11) is $\tilde{\theta}_i = p_i X_i$, where $p_i = \sum_{I: i\in I} \tilde{\pi}(I|X)$, and the estimator (2.13) is $\check{\theta}_i = X_i 1\{|X_i| \ge \check{t}\}$, where $\check{t} = |X_{[\check{k}]}|$, $|X_{[1]}| \ge \ldots \ge |X_{[n]}|$, and $\check{k}$ is the minimizer of $\sum_{i=k+1}^n X_{[i]}^2 + (2\kappa+1)\sigma^2 k\log(en/k)$.

The thresholding procedure is easy to implement, whereas the values p_i in the shrinkage procedure are more difficult to compute. It is demonstrated in [33] how one can use the partial product structure (in the model and in π_I, but not in λ_I) to facilitate the computation of p_i's. Other estimators can be considered, for example, the coordinatewise median with respect to $\tilde{\pi}$, which is going to be something in between shrinkage and thresholding.

Relaxing Condition (B0). We should mention that all the results still hold, if, instead of Condition (B0), we assume the weaker condition: $\mathbb{E}_\theta \exp\{\alpha \sum_{i\in I} \xi_i^2\} \le C_\alpha e^{|I|\log(en/|I|)}$ for

all $I \in \mathcal{I}$ and some $\alpha \in (0,1]$, $C_\alpha > 0$. However, we leave Condition (B0) in its present form to provide a cleaner mathematical exposition.

Relation to paper [86]. Recently the paper [86] on the same topic appeared (with discussion, see also our contribution [15] to this discussion). The main result of [86] is the adaptivity of the confidence set constructed by the Bayesian approach over the sparsity scale of nearly black vectors (introduced in Section 2.2.5) within a grand space $\ell_0[p_n]$ for some $p_n \to \infty$, $p_n = o(n)$ as $n \to \infty$, under the EBR condition. The EBR condition introduced in [86] is essentially a version of our EBR condition adopted to the sparsity scale within the grand space $\ell_0[p_n]$. It is not difficult to see that, within the asymptotic framework $n \to \infty$ and restricting the values of θ to some grand space $\ell_0[p_n]$ with $p_n = o(n)$, the EBR condition introduced in [86] is actually equivalent to our EBR condition specified to that embedded sparsity scale with appropriate choices of the constants involved.

Restricting the values of θ to some grand space $\ell_0[p_n]$ excludes some "almost sparse" parameters that are formally non-sparse (with many very small, but nonzero, entries), but this is in fact necessary to ensure the asymptotic regime $n \to \infty$ considered in [86]. The main differences of our approach and that of [86] are the following. We obtain local results without relating to any sparsity scale, e.g., the true parameter θ may be not $\ell_0[p_n]$-sparse at all. For example, as a consequence we derive the results not only for $\ell_0[p_n]$, but also for other sparsity scales, such as *weak* ℓ_q*-balls* $m_q[p_n]$. Next, we allow the error vector ξ to be non-normal and even not necessarily independent (but just satisfying Condition (B0)). Some of our constants in the proofs and conditions may be more conservative, which is not surprising since we pursue a more general situation. Finally, we derive non-asymptotic exponential concentration bounds, which give a refined characterization of the quality of coverage and size relation results (finer, than, e.g., Theorem 5 from [86], which is asymptotic in $n \to \infty$) and allow subtle analysis for various asymptotic regimes.

We should mention that the derivation of our somewhat stronger results relies on certain explicit posterior expressions resulting from our choice of prior and the used likelihood (although the model itself is not assumed to be normal), whereas the horseshoe prior studied in [86] leads to only implicit posterior quantities so that the authors had to overcome difficult technical issues in the proofs.

2.6. TECHNICAL LEMMAS

First we provide a couple of technical lemmas used in the proofs of the main results.

Remark 2.7. *Notice that in the below lemma we established the same bound for the both quantities* $\mathbb{E}_\theta \hat{\pi}(I|X) = \mathbb{E}_\theta \tilde{\pi}(I|X)$ *and* $\mathbb{E}_\theta \mathbb{1}\{\hat{I} = I\} = \mathbb{P}_\theta(\hat{I} = I)$. *The proofs of the properties of* $\check{\pi}(\vartheta|X)$ *and* $\check{\theta}$ *are exactly the same for* $\tilde{\pi}(\vartheta|X)$ *and* $\tilde{\theta}$, *with the only difference that everywhere (in the claims and in the proofs)* $\hat{\pi}(I \in \mathcal{G}|X)$ *should be read as* $\tilde{\pi}(I \in \mathcal{G}|X)$ *in case* $\hat{\pi} = \tilde{\pi}$; *and as* $\mathbb{1}\{\hat{I} \in \mathcal{G}\}$ *in case* $\hat{\pi} = \check{\pi}$, *for all* $\mathcal{G} \subseteq \mathcal{I}$ *that appear in the proof. Hence,* $\mathbb{E}_\theta \hat{\pi}(I \in \mathcal{G}|X) = \mathbb{E}_\theta \tilde{\pi}(I \in \mathcal{G}|X)$ *in the former case, and* $\mathbb{E}_\theta \hat{\pi}(I \in \mathcal{G}|X) = \mathbb{P}_\theta(\hat{I} \in \mathcal{G})$ *in the latter case.*

Lemma 2.1. *Let Condition (B0) be fulfilled. Then for any $\theta \in \mathbb{R}^n$ and any $I, I_0 \in \mathcal{I}$,*

$$\mathbb{E}_\theta \hat{\pi}(I|X) \le \big[\tfrac{\lambda_I}{\lambda_{I_0}}\big]^h \exp\Big\{B_h \sum_{i\in I\setminus I_0} \tfrac{\theta_i^2}{\sigma^2} - A_h \sum_{i\in I_0\setminus I} \tfrac{\theta_i^2}{\sigma^2} + C_h|I_0|\log(\tfrac{en}{|I_0|}) - D_h|I|\log(\tfrac{en}{|I|})\Big\},$$

where $h = \frac{2\alpha}{3}$, $A_h = \frac{\alpha}{6}$, $B_h = \frac{2\alpha}{3}$, $C_h = \frac{\alpha+1}{3}$ and $D_h = \frac{\alpha-2}{3}$. If $I\setminus I_0 = \varnothing$, the bound holds also for $h = \alpha$ with $A_h = \frac{\alpha}{3}$, $B_h = 0$, $C_h = \frac{\alpha}{2}+1$, $D_h = \frac{\alpha}{2}$. If $I_0\setminus I = \varnothing$, the bound holds also for $h = \alpha$ with $A_h = 0$, $B_h = \alpha$, $C_h = \frac{\alpha}{2}$, $D_h = \frac{\alpha}{2}-1$.

Proof of Lemma 2.1. Recall that $\mathbb{P}_{X,I} = \phi\big(X_i 1\{i \notin I\}, 0, \sigma^2 + K_n(I)\sigma^2 1\{i \in I\}\big)$. In case $\hat{\pi}(I|X) = \tilde{\pi}(I|X)$, we get by (2.10) that, for any $I, I_0 \in \mathcal{I}$ and any $h \in [0,1]$,

$$\begin{aligned}\mathbb{E}_\theta \hat{\pi}(I|X) = \mathbb{E}_\theta \tilde{\pi}(I|X) &= \mathbb{E}_\theta \frac{\lambda_I \mathbb{P}_{X,I}}{\sum_{J\in\mathcal{I}} \lambda_J \mathbb{P}_{X,J}} \le \mathbb{E}_\theta\Big(\frac{\lambda_I \mathbb{P}_{X,I}}{\lambda_{I_0}\mathbb{P}_{X,I_0}}\Big)^h \qquad (2.30)\\ &= \mathbb{E}_\theta\Big[\frac{\lambda_I \prod_{i=1}^n \phi\big(X_i 1\{i\notin I\}, 0, \sigma^2 + K_n(I)\sigma^2 1\{i\in I\}\big)}{\lambda_{I_0}\prod_{i=1}^n \phi\big(X_i 1\{i\notin I_0\}, 0, \sigma^2 + K_n(I_0)\sigma^2 1\{i\in I_0\}\big)}\Big]^h \\ &= \big[\tfrac{\lambda_I}{\lambda_{I_0}}\big]^h \mathbb{E}_\theta \exp\Big\{\frac{h}{2}\Big[\sum_{i\in I\setminus I_0}\frac{X_i^2}{\sigma^2} - \sum_{i\in I_0\setminus I}\frac{X_i^2}{\sigma^2} + |I_0|\log(\tfrac{en}{|I_0|}) - |I|\log(\tfrac{en}{|I|})\Big]\Big\}. \qquad (2.31)\end{aligned}$$

In case $\hat{\pi}(I|X) = \mathbb{1}\{\hat{I} = I\}$, by the definition (2.12) of $\hat{I}$ and the Markov inequality, we derive that, for any $I, I_0 \in \mathcal{I}$ and any $h \ge 0$,

$$\mathbb{E}_\theta \hat{\pi}(I|X) = \mathbb{P}_\theta(\hat{I} = I) \le \mathbb{P}_\theta\Big(\frac{\tilde{\pi}(I|X)}{\tilde{\pi}(I_0|X)} \ge 1\Big) \le \mathbb{E}_\theta\Big[\frac{\tilde{\pi}(I|X)}{\tilde{\pi}(I_0|X)}\Big]^h = \mathbb{E}_\theta\Big(\frac{\lambda_I \mathbb{P}_{X,I}}{\lambda_{I_0}\mathbb{P}_{X,I_0}}\Big)^h,$$

which yields exactly the bound (2.30), and hence the bound (2.31) again.

Using Hölder's inequality, Condition (B0) and the two elementary facts $X_i^2 \le 2\theta_i^2 + 2\sigma^2\xi_i^2$ and $-X_i^2 \le -\frac{\theta_i^2}{2} + \sigma^2\xi_i^2$, we obtain

$$\begin{aligned}\mathbb{E}_\theta \exp\Big\{\frac{\alpha}{3}\Big[\sum_{i\in I\setminus I_0}\frac{X_i^2}{\sigma^2} - \sum_{i\in I_0\setminus I}\frac{X_i^2}{\sigma^2}\Big]\Big\} &\le \Big(\mathbb{E}_\theta e^{\frac{\alpha}{2}\sum_{i\in I\setminus I_0}\frac{X_i^2}{\sigma^2}}\Big)^{2/3}\Big(\mathbb{E}_\theta e^{-\alpha\sum_{i\in I_0\setminus I}\frac{X_i^2}{\sigma^2}}\Big)^{1/3} \\ &\le \exp\Big\{\frac{2\alpha}{3}\sum_{i\in I\setminus I_0}\frac{\theta_i^2}{\sigma^2} + \frac{2}{3}|I\setminus I_0| - \frac{\alpha}{6}\sum_{i\in I_0\setminus I}\frac{\theta_i^2}{\sigma^2} + \frac{1}{3}|I_0\setminus I|\Big\}.\end{aligned}$$

Since $|I\setminus I_0| \le |I| \le |I|\log(\frac{en}{|I|})$ and $|I_0\setminus I| \le |I_0| \le |I_0|\log(\frac{en}{|I_0|})$, the lemma follows for $h = \frac{2\alpha}{3}$ from the last display and (2.31).

If $I\setminus I_0 = \varnothing$, we take $h = \alpha$ in (2.31) and combine this with $\mathbb{E}_\theta \exp\big\{-\frac{\alpha}{2}\sum_{i\in I_0\setminus I}\frac{X_i^2}{\sigma^2}\big\} \le \exp\big\{-\frac{\alpha}{3}\sum_{i\in I_0\setminus I}\frac{\theta_i^2}{\sigma^2} + |I_0\setminus I|\big\}$, which holds in view of Condition (B0) and $-\frac{X_i^2}{\sigma^2} \le -\frac{2\theta_i^2}{3\sigma^2} + 2\xi_i^2$, as $(a+b)^2 \ge 2a^2/3 - 2b^2$. If $I_0\setminus I = \varnothing$, we take $h = \alpha$ in (2.31) and combine this with $\mathbb{E}_\theta \exp\big\{\frac{\alpha}{2}\sum_{i\in I\setminus I_0}\frac{X_i^2}{\sigma^2}\big\} \le \exp\big\{\alpha\sum_{i\in I\setminus I_0}\frac{\theta_i^2}{\sigma^2} + |I\setminus I_0|\big\}$ which holds in view of Condition (B0) and $\frac{X_i^2}{\sigma^2} \le \frac{2\theta_i^2}{\sigma^2} + 2\xi_i^2$. □

Note that above lemma holds for any set $I_0 \in \mathcal{I}$. By taking $I_0 = I_o$ defined by (2.17), we obtain the following lemma.

Lemma 2.2. *Let Conditions* (B0) *and* (B1) *be fulfilled. Then there exist positive constants* $c_1 = c_1(\kappa) > 2$, c_2 *and* $c_3 = c_3(\kappa)$ *such that for any* $\theta \in \mathbb{R}^n$

$$\mathbb{E}_\theta \hat{\pi}(I|X) \le \bigl(\tfrac{ne}{|I|}\bigr)^{-c_1|I|} \exp\bigl\{-c_2\sigma^{-2}\bigl[r^2(I,\theta) - c_3 r^2(\theta)\bigr]\bigr\}.$$

Proof of Lemma 2.2. With constants h, A_h, B_h, C_h, D_h given in Lemma 2.1, define the constant $c_1 = c_1(\kappa) = \kappa h + D_h - A_h = \frac{2\alpha\kappa}{3} + \frac{\alpha-2}{3} - \frac{\alpha}{6} > 2$ as $\kappa > \bar{\kappa}$ by Condition (B1). Since $\kappa h + D_h = c_1 + A_h$, the definition (2.4) of λ_I entails that

$$\begin{aligned}
&\bigl(\tfrac{\lambda_I}{\lambda_{I_0}}\bigr)^h \exp\bigl\{C_h|I_0|\log(\tfrac{en}{|I_0|}) - D_h|I|\log(\tfrac{en}{|I|})\bigr\} \\
&= (\tfrac{ne}{|I|})^{-c_1|I|} \exp\bigl\{(\kappa h + C_h)|I_0|\log(\tfrac{en}{|I_0|}) - A_h|I|\log(\tfrac{en}{|I|})\bigr\}.
\end{aligned}$$

Using the last relation and Lemma 2.1 with $I_0 = I_o$, we bound

$$\begin{aligned}
&\mathbb{E}_\theta \hat{\pi}(I|X) \le \bigl[\tfrac{\lambda_I}{\lambda_{I_o}}\bigr]^h \exp\Bigl\{B_h \sum_{i\in I\setminus I_o} \frac{\theta_i^2}{\sigma^2} - A_h \sum_{i\in I_o\setminus I} \frac{\theta_i^2}{\sigma^2} + C_h|I_o|\log(\tfrac{en}{|I_o|}) - D_h|I|\log(\tfrac{en}{|I|})\Bigr\} \\
&= (\tfrac{ne}{|I|})^{-c_1|I|} \exp\Bigl\{-A_h \sum_{i\in I_o\setminus I} \frac{\theta_i^2}{\sigma^2} - A_h|I|\log(\tfrac{en}{|I|}) + B_h \sum_{i\in I\setminus I_o} \frac{\theta_i^2}{\sigma^2} + (\kappa h + C_h)|I_o|\log(\tfrac{en}{|I_o|})\Bigr\}.
\end{aligned}$$

The claim of the lemma follows with the constants $c_1 = (4\alpha\kappa + \alpha - 4)/6 > 2$, $c_2 = A_h = \alpha/6$ and $c_3 = c_3(\kappa) = \max\{B_h, \kappa h + C_h\}/A_h = (\kappa h + C_h)/A_h = 4\kappa + 2(\alpha+1)/\alpha$. □

Lemma 2.3. *Let* $Y_1, \ldots, Y_n$ *be some random variables such that, for any* $I \in \mathcal{I}$, $\mathbb{E}_\theta e^{t\sum_{i\in I} Y_i} \le A_{|I|}(t)$ *for some* $t > 0$ *and* $A_k(t)$. *Let* $Y_{[1]} \ge Y_{[2]} \ge \ldots \ge Y_{[n]}$. *Then, for any* $k \in \mathbb{N}_n$ *and* $C, c \ge 0$,

$$\begin{aligned}
&\mathbb{P}_\theta\Bigl(\sum_{i=1}^k Y_{[i]} \ge Ck\log(\tfrac{en}{k}) + c\Bigr) \le A_k(t)\exp\{-(Ct-1)k\log(\tfrac{en}{k}) - ct\}, \\
&\mathbb{E}_\theta \sum_{i=1}^k Y_{[i]} \le t^{-1}\bigl[k\log(\tfrac{en}{k}) + \log(A_k(t))\bigr].
\end{aligned}$$

In particular, if $\xi_1, \ldots, \xi_n \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, *then for any* $k \in \mathbb{N}_n$, $C, c \ge 0$

$$\mathbb{P}_\theta\Bigl(\sum_{i=1}^k \xi_{[i]}^2 \ge Ck\log\bigl(\tfrac{en}{k}\bigr) + c\Bigr) \le \bigl(\tfrac{en}{k}\bigr)^{-(0.4C-2)k} e^{-0.4c}, \quad \mathbb{E}_\theta \sum_{i=1}^k \xi_{[i]}^2 \le 6k\log\bigl(\tfrac{en}{k}\bigr).$$

Proof. By Jensen's inequality, we derive

$$\exp\Bigl\{t\mathbb{E}_\theta \sum_{i=1}^k Y_{[i]}\Bigr\} \le \mathbb{E}_\theta \exp\Bigl\{t\sum_{i=1}^k Y_{[i]}\Bigr\} \le \sum_{I:|I|=k} \mathbb{E}_\theta \exp\Bigl\{t\sum_{i\in I} Y_i\Bigr\} \le \tbinom{n}{k} A_k(t).$$

Then $\mathbb{E}_\theta \exp\{t\sum_{i=1}^k Y_{[i]}\} \le \binom{n}{k} A_k(t) \le e^{k\log(\frac{en}{k}) + \log(A_k(t))}$, where we used $\binom{n}{k} \le (\frac{en}{k})^k$. This and the (exponential) Markov inequality yield the first relation:

$$\mathbb{P}_\theta\Bigl(\sum_{i=1}^k Y_{[i]} \ge Ck\log\bigl(\frac{en}{k}\bigr) + c\Bigr) \le A_k(t)\exp\{-(Ct-1)k\log(\tfrac{en}{k}) - ct\}.$$

The first display implies also the second relation: $\mathbb{E}_\theta \sum_{i=1}^k Y_{[i]} \le t^{-1}[\log\binom{n}{k} + \log(A_k(t))]$.

As to the normal case, for any $I \in \mathcal{I}$ and any $t < \frac{1}{2}$ we have that $\mathbb{E}_\theta \exp\{t\sum_{i\in I}\xi_i^2\} = (1-2t)^{-|I|/2} = A_{|I|}(t)$. Since $A_k(t) \le e^k \le e^{k\log(\frac{en}{k})}$ for any $t \le (1-e^{-2})/2 < 0.43$, the first assertion for the normal case follows by taking $t = 0.4$. By taking $t = \frac{1}{4}$, the second assertion follows since $\mathbb{E}_\theta \sum_{i=1}^k \xi_{[i]}^2 \le 4k\log(\frac{en}{k}) + 2k\log 2 \le 6k\log(\frac{en}{k})$. □

This lemma is useful if $A_k(t) \le C_1(\frac{en}{k})^{C_2 k}$ for some $t, C_1, C_2 > 0$; in particular, for $Y_i = \xi_i^2$, where the ξ_i's satisfy Condition (B0). Then Lemma 2.3 applies with $t = \alpha$ and $A_k(\alpha) = e^k$:

$$\mathbb{P}_\theta\Big(\sum_{i=1}^k \xi_{[i]}^2 \ge \tfrac{2}{\alpha} k\log(\tfrac{en}{k}) + M\Big) \le \exp\{-\alpha M\}, \quad k \in \mathbb{N}_n,\ M \ge 0. \tag{2.32}$$

2.7. PROOFS OF THE THEOREMS

Here we gather the proofs of the theorems. By C_0, C_1, C_2 etc., denote constants which are different in different proofs. Recall that $Y_{[1]} \ge Y_{[2]} \ge \ldots \ge Y_{[n]}$ denote the ordered $Y_1, \ldots, Y_n$.

Proof of Theorem 2.1. Recall the constants c_1, c_2, c_3 defined in the proof of Lemma 2.2. Let $M_0 = 2c_3(6+\frac{2}{\alpha})$. Introduce the subfamily of index sets $\mathcal{S}_M = \mathcal{S}_M(\theta) = \{I \in \mathcal{I} : r^2(I,\theta) \le c_3 r^2(\theta) + \frac{\alpha}{80} M\sigma^2\}$, $m = m_M(\theta) = \max\{|I| : I \in \mathcal{S}_M\}$, and the event $A_M = A(\theta) = \{\sum_{i=1}^m \xi_{[i]}^2 \le \frac{2}{\alpha} m\log(\frac{en}{m}) + \frac{M}{8}\}$. We have

$$\begin{aligned}\hat\pi\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2 | X\big) &\le \mathbb{1}_{A_M^c} + \hat\pi(I \in \mathcal{S}_M^c | X) \\ &+ \sum_{I\in\mathcal{S}_M} \mathbb{1}_{A_M} \hat\pi_I\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2 | X\big)\hat\pi(I|X) = T_1 + T_2 + T_3.\end{aligned}$$

Now we bound the quantities $\mathbb{E}_\theta T_1$, $\mathbb{E}_\theta T_2$ and $\mathbb{E}_\theta T_3$.

First, we bound $\mathbb{E}_\theta T_1$ by using Lemma 2.3 (see also (2.32)):

$$\mathbb{E}_\theta T_1 = \mathbb{P}_\theta(A_M^c) = \mathbb{P}_\theta\Big(\sum_{i=1}^m \xi_{[i]}^2 > \tfrac{2}{\alpha} m\log(\tfrac{en}{m}) + \tfrac{M}{8}\Big) \le \exp\big\{-\alpha M/8\big\}. \tag{2.33}$$

Let us bound $\mathbb{E}_\theta T_2$. Since $\binom{n}{k} \le (\frac{en}{k})^k$ and $c_1 > 2$, the following relation holds:

$$\sum_{I\in\mathcal{I}} \big(\tfrac{ne}{|I|}\big)^{-c_1|I|} = \sum_{k=0}^n \binom{n}{k}\big(\tfrac{en}{k}\big)^{-c_1 k} \le \sum_{k=0}^n \big(\tfrac{en}{k}\big)^{-k(c_1-1)} \le (1-e^{1-c_1})^{-1} \triangleq C_0. \tag{2.34}$$

If $I \in \mathcal{S}_M^c$, then $r^2(I,\theta) > c_3 r^2(\theta) + \frac{\alpha}{80} M\sigma^2$. Using this, Lemma 2.2 and (2.34), we bound $\mathbb{E}_\theta T_2$:

$$\begin{aligned}\mathbb{E}_\theta T_2 = \sum_{I\in\mathcal{S}_M^c} \mathbb{E}_\theta \hat\pi(I|X) &\le \sum_{I\in\mathcal{S}_M^c} \big(\tfrac{ne}{|I|}\big)^{-c_1|I|} \exp\big\{-c_2\sigma^{-2}\big[r^2(I,\theta) - c_3 r^2(\theta)\big]\big\} \\ &\le \sum_{I\in\mathcal{I}} \big(\tfrac{ne}{|I|}\big)^{-c_1|I|} \exp\{-c_2\alpha M/80\} \le C_0 \exp\{-c_2\alpha M/80\}.\end{aligned} \tag{2.35}$$

It remains to bound $\mathbb{E}_\theta T_3$. For each $I \in \mathcal{S}_M$, $\sigma^2|I|\log(en/|I|) \le r^2(I,\theta) \le c_3 r^2(\theta) + \frac{\alpha}{80}M\sigma^2$. Since $m = \max\{|I| : I \in \mathcal{S}_M\}$, then $\sigma^2 m \log(\frac{en}{m}) \le c_3 r^2(\theta) + \frac{\alpha}{80}M\sigma^2$. Thus, for any $I \in \mathcal{S}_M$, the event A_M implies that $\sum_{i\in I}\xi_i^2 \le \sum_{i=1}^m \xi_{[i]}^2 \le \frac{2}{\alpha} m\log(\frac{en}{m}) + \frac{M}{8} \le \frac{2}{\alpha}c_3\sigma^{-2}r^2(\theta) + \frac{3M}{20}$. Denote for brevity $\Delta_M(\theta) = M_0 r^2(\theta) + M\sigma^2$ and recall that $\sum_{i\in I^c}\theta_i^2 \le r^2(I,\theta) \le c_3 r^2(\theta) + \frac{\alpha}{80}M\sigma^2 \le c_3 r^2(\theta) + \frac{M}{80}\sigma^2$ for any $I \in \mathcal{S}_M$. Then for any $I \in \mathcal{S}_M$

$$A_M \subseteq \Big\{\frac{\Delta_M(\theta)}{2} - \sigma^2\sum_{i\in I}\xi_i^2 - \sum_{i\in I^c}\theta_i^2 \ge \Big[\frac{M_0}{2} - \frac{2+\alpha}{\alpha}c_3\Big]r^2(\theta) + \frac{27M\sigma^2}{80}\Big\}. \tag{2.36}$$

According to (2.9), $\hat\pi_I(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}\big(X_i(I), \sigma_i^2(I)\big)$, with $X_i(I) = X_i 1\{i \in I\}$ and $\sigma_i^2(I) = \frac{K_n(I)\sigma^2 1\{i\in I\}}{K_n(I)+1}$. Let $\mathbb{P}_Z$ be the measure of $Z = (Z_1,\dots,Z_n)$, with $Z_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. By using (2.36), the fact that $\frac{r^2(\theta)}{\sigma^2} \ge c_3^{-1}(m\log(\frac{en}{m}) - \frac{\alpha}{80}M)$ and Lemma 2.3 (now applied to the Gaussian case), we obtain that, for any $I \in \mathcal{S}_M$,

$$\begin{aligned}
\hat\pi_I\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2|X\big)\mathbb{1}_{A_M} &= \mathbb{P}_Z\Big(\sum_{i=1}^n(\sigma_i(I)Z_i + X_i(I) - \theta_i)^2 \ge \Delta_M(\theta)\Big)\mathbb{1}_{A_M}\\
&\le \mathbb{P}_Z\Big(\sum_{i=1}^n \sigma_i^2(I)Z_i^2 \ge \tfrac{\Delta_M(\theta)}{2} - \sum_{i=1}^n (X_i(I)-\theta_i)^2\Big)\mathbb{1}_{A_M}\\
&\le \mathbb{P}_Z\Big(\sum_{i\in I}\sigma^2 Z_i^2 \ge \tfrac{\Delta_M(\theta)}{2} - \sum_{i\in I}\sigma^2\xi_i^2 - \sum_{i\in I^c}\theta_i^2\Big)\mathbb{1}_{A_M}\\
&\le \mathbb{P}_Z\Big(\sum_{i\in I} Z_i^2 \ge \big[\tfrac{M_0}{2} - \big(\tfrac{2}{\alpha}+1\big)c_3\big]\frac{r^2(\theta)}{\sigma^2} + \tfrac{27M}{80}\Big)\\
&\le \mathbb{P}_Z\Big(\sum_{i=1}^m Z_{[i]}^2 \ge \big(\tfrac{M_0}{2c_3} - \tfrac{2}{\alpha} - 1\big)\big[m\log(\tfrac{en}{m}) - \tfrac{\alpha}{80}M\big] + \tfrac{27M}{80}\Big)\\
&\le \mathbb{P}_Z\Big(\sum_{i=1}^m Z_{[i]}^2 \ge 5m\log(\tfrac{en}{m}) + \tfrac{11M}{40}\Big) \le \exp\{-11M/100\},
\end{aligned}$$

where we also used in the last step that $\frac{M_0}{2c_3} - \frac{2}{\alpha} - 1 = 5$. Hence,

$$\begin{aligned}
\mathbb{E}_\theta T_3 &= \mathbb{E}_\theta \sum_{I\in\mathcal{S}_M} \mathbb{1}_{A_M}\hat\pi_I\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2|X\big)\hat\pi(I|X)\\
&\le \exp\{-11M/100\}\mathbb{E}_\theta\sum_{I\in\mathcal{I}}\hat\pi(I|X) \le \exp\{-11M/100\}.
\end{aligned}$$

This completes the proof of assertion (i) since, in view of (2.33), (2.35) and the last display, we established that $\mathbb{E}_\theta\hat\pi\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2|X\big) \le \mathbb{E}_\theta(T_1+T_2+T_3) \le (2+C_0)e^{-m_0 M}$, with constants $M_0 = 2c_3(6+\frac{2}{\alpha})$, $H_0 = 2 + C_0$, $m_0 = \min\{\frac{\alpha}{8}, \frac{c_2\alpha}{80}, \frac{11}{100}\}$ and C_0 defined in (2.34).

The proof of assertion (ii) proceeds along similar lines. Recall the constants $c_1 > 2$, c_2, c_3 from Lemma 2.2 and define $M_1 = 4c_3(2+\alpha)/\alpha$. Introduce the subfamily of sets

$$\bar{\mathcal{S}}_M = \bar{\mathcal{S}}_M(\theta) = \big\{I \in \mathcal{I} : r^2(I,\theta) \le 2c_3 r^2(\theta) + \tfrac{\alpha}{12}M\sigma^2\big\},$$

and the event $\bar A_M = \bar A_M(\theta) = \big\{\sum_{i=1}^{\bar m}\xi_{[i]}^2 \le \frac{2}{\alpha}\bar m\log(\frac{en}{\bar m}) + \frac{M}{6}\big\}$, where $\bar m = \bar m_M(\theta) = \max\{|I| : I \in \bar{\mathcal{S}}_M\}$. Introduce the notation $\bar\Delta_M(\theta) = M_1 r^2(\theta) + M\sigma^2$ for brevity. By the definition of

$\hat\theta$ and the Cauchy-Schwarz inequality, we have that $\|\hat\theta-\theta\|^2 \le \sum_{I\in\mathcal{I}} \|X(I)-\theta\|^2 \hat\pi(I|X)$, where $\|X(I)-\theta\|^2 = \sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2$. Using this, we derive

$$\begin{aligned}
&\mathbb{P}_\theta\big(\|\hat\theta-\theta\|^2 \ge \bar\Delta_M(\theta)\big) \le \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{I}} \|X(I)-\theta\|^2\hat\pi(I|X) \ge \bar\Delta_M(\theta)\Big)\\
&\le \mathbb{P}_\theta(\bar A_M^c) + \mathbb{P}_\theta\Big(\Big\{\sum_{I\in\bar{\mathcal{S}}_M}\Big[\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\Big]\hat\pi(I|X) \ge \bar\Delta_M(\theta)/2\Big\}\cap \bar A_M\Big)\\
&\quad + \mathbb{P}_\theta\Big(\sum_{I\in\bar{\mathcal{S}}_M^c}\Big[\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\Big]\hat\pi(I|X) \ge \bar\Delta_M(\theta)/2\Big) = \bar T_1 + \bar T_2 + \bar T_3.
\end{aligned}$$

Similar to (2.33), we bound the term $\bar T_1$ by Lemma 2.3 (see also (2.32)):

$$\bar T_1 = \mathbb{P}_\theta(\bar A_M^c) = \mathbb{P}_\theta\Big(\sum_{i=1}^{\bar m}\xi_{[i]}^2 > \tfrac{2}{\alpha}\bar m\log(\tfrac{en}{\bar m}) + \tfrac{M}{6}\Big) \le \exp\big\{-M\alpha/6\big\}.$$

Now we evaluate the term $\bar T_2$. Since $\bar m = \max\{|I| : I\in\bar{\mathcal{S}}_M\}$, $\sigma^2\bar m\log(\frac{en}{\bar m}) \le 2c_3r^2(\theta) + \frac{\alpha}{12}M\sigma^2$. Then for any $I\in\bar{\mathcal{S}}_M$, the event $\bar A_M$ implies that $\sum_{i\in I}\xi_i^2 \le \sum_{i=1}^{\bar m}\xi_{[i]}^2 \le \frac{2}{\alpha}\bar m\log(\frac{en}{\bar m}) + \frac{M}{6} \le \frac{4c_3r^2(\theta)}{\alpha\sigma^2} + \frac{M}{3}$. Also $\sum_{i\in I^c}\theta_i^2 \le r^2(I,\theta) \le 2c_3r^2(\theta) + \frac{\alpha}{12}M\sigma^2$ for any $I\in\bar{\mathcal{S}}_M$. Hence, for any $I\in\bar{\mathcal{S}}_M$, we obtain the implication

$$\bar A_M \subseteq \Big\{\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2 \le \tfrac{2c_3(2+\alpha)}{\alpha}r^2(\theta) + (\tfrac13+\tfrac{\alpha}{12})M\sigma^2\Big\}.$$

As $M_1 = 4c_3(2+\alpha)/\alpha$, $\alpha\in(0,1]$, the last relation entails

$$\begin{aligned}
\bar T_2 &= \mathbb{P}_\theta\Big(\Big\{\sum_{I\in\bar{\mathcal{S}}_M}\Big(\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\Big)\hat\pi(I|X) \ge \tfrac{\bar\Delta_M}{2}\Big\}\cap\bar A_M\Big)\\
&\le \mathbb{P}_\theta\Big(\tfrac{2c_3(2+\alpha)}{\alpha}r^2(\theta) + (\tfrac13+\tfrac{\alpha}{12})M\sigma^2 \ge \tfrac{M_1}{2}r^2(\theta) + \tfrac{M}{2}\sigma^2\Big) = 0.
\end{aligned}$$

It remains to handle the term $\bar T_3$. Applying first the Markov inequality and then the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\bar T_3 &\le \frac{\mathbb{E}_\theta\big(\sum_{I\in\bar{\mathcal{S}}_M^c}\big[\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\big]\hat\pi(I|X)\big)}{\bar\Delta_M(\theta)/2}\\
&\le \frac{\sum_{I\in\bar{\mathcal{S}}_M^c}\big(\sigma^2\big[\mathbb{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2\big]^{1/2}\big[\mathbb{E}_\theta(\hat\pi(I|X))^2\big]^{1/2} + r^2(I,\theta)\mathbb{E}_\theta\hat\pi(I|X)\big)}{\bar\Delta_M(\theta)/2} = T_{31}+T_{32}.
\end{aligned}$$

For any $I\in\bar{\mathcal{S}}_M^c$, we have $c_3r^2(\theta) \le \frac{r^2(I,\theta)}{2} - \frac{\alpha}{24}M\sigma^2$, yielding the bound

$$\frac{c_2}{2}\big(r^2(I,\theta) - c_3r^2(\theta)\big) \ge C_1r^2(I,\theta) + C_2M\sigma^2 \quad \text{for any } I\in\bar{\mathcal{S}}_M^c, \tag{2.37}$$

where $C_1 = c_2/4$ and $C_2 = c_2\alpha/48$. By (2.37) and Lemma 2.2,

$$\big[\mathbb{E}_\theta\hat\pi(I|X)\big]^{1/2} \le \big(\tfrac{ne}{|I|}\big)^{-c_1|I|/2}\exp\big\{-C_1\sigma^{-2}r^2(I,\theta) - C_2M\big\} \quad \text{for any } I\in\bar{\mathcal{S}}_M^c. \tag{2.38}$$

Since $c_1 > 2$, (2.34) gives $\sum_{I\in\mathcal{I}}\big(\frac{ne}{|I|}\big)^{-c_1|I|/2} \le (1-e^{-c_1/2})^{-1} \triangleq C_3$. According to (2.15) with $\rho = \min\{C_1, 1/2\} = C_1$, $\Big[\mathbb{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2\Big]^{1/2} \le \frac{1}{\alpha\rho}\exp\{\rho|I|\}$. If $M \in [0,1]$, the claim (ii) holds for any $H_1 \ge e^{m_1}$. Let $M \ge 1$, then $\sigma^2/\bar{\Delta}_M(\theta) \le M^{-1} \le 1$. Besides, $\sigma^{-2}r^2(I,\theta) \ge |I|\log(en/|I|) \ge |I|$. Piecing all these relations together with (2.38), we derive

$$T_{31} \le \tfrac{2}{\alpha\rho}\sum_{I\in\bar{\mathcal{S}}_M^c}\exp\{\rho|I|\}\big(\tfrac{ne}{|I|}\big)^{-c_1|I|/2}\exp\big\{-C_1\sigma^{-2}r^2(I,\theta)-C_2M\big\} \le C_4\exp\{-C_2M\},$$

where $C_4 = 2C_3/(\alpha\rho) = 2C_3/(\alpha C_1)$. Finally, by (2.34), (2.38) and the facts that $\max_{x\ge0}\{xe^{-cx}\} \le (ce)^{-1}$ (for any $c>0$) and $\sigma^2/\bar{\Delta}_M(\theta) \le 1$, we bound the term T_{32}:

$$\begin{aligned} T_{32} &= \tfrac{2}{\bar{\Delta}_M(\theta)}\sum_{I\in\bar{\mathcal{S}}_M^c} r^2(I,\theta)\mathbb{E}_\theta\hat{\pi}(I|X) \\ &\le \tfrac{2}{\bar{\Delta}_M(\theta)}\sum_{I\in\bar{\mathcal{S}}_M^c} r^2(I,\theta)\big(\tfrac{ne}{|I|}\big)^{-c_1|I|}\exp\big\{-2C_1\sigma^{-2}r^2(I,\theta)-2C_2M\big\} \le C_5\exp\{-2C_2M\}, \end{aligned}$$

where $C_5 = C_0/(C_1e)$. The assertion (ii) is proved since we showed that $\mathbb{P}_\theta\big(\|\hat{\theta}-\theta\|^2 \ge M_1r^2(\theta)+M\sigma^2\big) \le H_1e^{-m_1M}$ with $M_1 = 4c_3(2+\alpha)/\alpha$, $H_1 = (1+C_4+C_5)\vee e^{m_1}$, $m_1 = \min\{\frac{\alpha}{6}, C_2\}$. □

Proof of Theorem 2.2. First we prove (i). If the inequality $|I\setminus I_o|\log(\frac{en}{|I|}) < \sum_{i\in I\setminus I_o}\frac{\theta_i^2}{\sigma^2}$ would hold for some $I\in\mathcal{I}$, then

$$\begin{aligned} r^2(I\cup I_o,\theta) &= \sum_{i\notin I\cup I_o}\theta_i^2 + \sigma^2|I\cup I_o|\log(\tfrac{en}{|I\cup I_o|}) \le \sum_{i\notin I\cup I_o}\theta_i^2 + \sigma^2|I\setminus I_o|\log(\tfrac{en}{|I|}) + \sigma^2|I_o|\log(\tfrac{en}{|I_o|}) \\ &< \sum_{i\notin I\cup I_o}\theta_i^2 + \sum_{i\in I\setminus I_o}\tfrac{\theta_i^2}{\sigma^2} + \sigma^2|I_o|\log(\tfrac{en}{|I_o|}) = \sum_{i\notin I_o}\theta_i^2 + \sigma^2|I_o|\log(\tfrac{en}{|I_o|}) = r^2(\theta), \end{aligned}$$

which contradicts the definition of the oracle. Hence, $\sum_{i\in I\setminus I_o}\frac{\theta_i^2}{\sigma^2} \le |I\setminus I_o|\log(\frac{en}{|I|})$ for any $I\in\mathcal{I}$. Define $c_4 = \kappa\alpha - \frac{\alpha}{2} - 2$ and note that $c_4 > 1$ by the condition of the theorem. Using the relation $\sum_{i\in I\setminus I_o}\frac{\theta_i^2}{\sigma^2} \le |I\setminus I_o|\log(\frac{en}{|I|}) \le |I|\log(\frac{en}{|I|})$ and Lemma 2.1 with $h=\alpha$ and $I_0 = I_o\cap I$ (so that $I\setminus I_0 = I\setminus I_o$), we obtain for each $I\in\mathcal{G}_1 = \{I\in\mathcal{I}: |I|\log(\frac{en}{|I|}) \ge M_0'|I_0|\log(\frac{en}{|I_0|}) + M\}$ with $M_0' = \kappa\alpha + \frac{\alpha}{2}$,

$$\begin{aligned} \mathbb{E}_\theta\hat{\pi}(I|X) &\le \big[\tfrac{\lambda_I}{\lambda_{I_0}}\big]^\alpha\exp\Big\{\alpha\sum_{i\in I\setminus I_0}\tfrac{\theta_i^2}{\sigma^2} + \tfrac{\alpha}{2}|I_0|\log(\tfrac{en}{|I_0|}) - (\tfrac{\alpha}{2}-1)|I|\log(\tfrac{en}{|I|})\Big\} \\ &\le (\tfrac{ne}{|I|})^{-c_4|I|}\exp\big\{-(\kappa\alpha-\tfrac{\alpha}{2}-1-c_4)|I|\log(\tfrac{en}{|I|}) + (\alpha\kappa+\tfrac{\alpha}{2})|I_0|\log(\tfrac{en}{|I_0|})\big\} \\ &= (\tfrac{ne}{|I|})^{-c_4|I|}\exp\big\{-|I|\log(\tfrac{en}{|I|}) + (\alpha\kappa+\tfrac{\alpha}{2})|I_0|\log(\tfrac{en}{|I_0|})\big\} \\ &\le (\tfrac{ne}{|I|})^{-c_4|I|}\exp\big\{-(M_0'-\kappa\alpha-\tfrac{\alpha}{2})|I_0|\log(\tfrac{en}{|I_0|}) - M\big\} = (\tfrac{ne}{|I|})^{-c_4|I|}e^{-M}. \end{aligned}$$

Since $c_4 > 1$, by the same reasoning as in (2.5) we bound $\sum_{I\in\mathcal{I}}(\frac{ne}{|I|})^{-c_4|I|} \le (1-e^{1-c_4})^{-1} \triangleq H_0'$. Using this and the last display, we finish the proof of (i):

$$\mathbb{E}_\theta\hat{\pi}(I\in\mathcal{G}_1|X) = \sum_{I\in\mathcal{G}_1}\mathbb{E}_\theta\hat{\pi}(I|X) \le e^{-M}\sum_{I\in\mathcal{I}}(\tfrac{ne}{|I|})^{-c_4|I|} \le H_0'e^{-M}.$$

Next we prove (ii). Define $\mathcal{G}_2 = \mathcal{G}_2(I') = \{I \in \mathcal{I} : \sum_{i\in I'\setminus I} \frac{\theta_i^2}{\sigma^2} \ge \bar\tau |I\cup I'| \log(\frac{en}{|I\cup I'|}) + M\}$. Using (2.4) and Lemma 2.1 with $h=\alpha$ and $I_0 = I_0(I,\theta) = I\cup I'$, we evaluate for each $I\in\mathcal{G}_2$

$$\begin{aligned}
\mathbb{E}_\theta\hat\pi(I|X) &\le \big[\tfrac{\lambda_I}{\lambda_{I_0}}\big]^\alpha \exp\Big\{-\tfrac{\alpha}{3}\sum_{i\in I_0\setminus I}\frac{\theta_i^2}{\sigma^2} + (\tfrac{\alpha}{2}+1)|I_0|\log\big(\tfrac{en}{|I_0|}\big) - \tfrac{\alpha}{2}|I|\log\big(\tfrac{en}{|I|}\big)\Big\}\\
&= \big[\tfrac{\lambda_I}{c_{\kappa,n}}\big]^\alpha \exp\Big\{-\tfrac{\alpha}{3}\sum_{i\in I'\setminus I}\frac{\theta_i^2}{\sigma^2} + (\kappa\alpha+\tfrac{\alpha}{2}+1)|I\cup I'|\log\big(\tfrac{en}{|I\cup I'|}\big) - \tfrac{\alpha}{2}|I|\log\big(\tfrac{en}{|I|}\big)\Big\}\\
&\le \big[\tfrac{\lambda_I}{c_{\kappa,n}}\big]^{\alpha+\frac{\alpha}{2\kappa}} \exp\big\{\big(-\tfrac{\alpha}{3}\bar\tau + \kappa\alpha + \tfrac{\alpha}{2}+1\big)|I\cup I'|\log\big(\tfrac{en}{|I\cup I'|}\big) - \tfrac{\alpha}{3}M\big\} \le \big[\tfrac{\lambda_I}{c_{\kappa,n}}\big]^{\alpha+\frac{\alpha}{2\kappa}} e^{-\frac{\alpha}{3}M}.
\end{aligned}$$

Since $\kappa > \alpha^{-1} - \frac12$, by the same reasoning as in (2.5) we bound $\sum_I \big[\frac{\lambda_I}{c_{\kappa,n}}\big]^{\alpha(1+1/2\kappa)} \le (1 - e^{1-\kappa\alpha-\alpha/2})^{-1} \triangleq H_1'$. This relation and the last display imply claim (ii): with $m_0' = \frac{\alpha}{3}$,

$$\mathbb{E}_\theta\hat\pi\big(I\in\mathcal{G}_2|X\big) = \sum_{I\in\mathcal{G}_2}\mathbb{E}_\theta\hat\pi(I|X) \le H_1'\exp\{-m_0'M\}. \tag{2.39}$$

Let us derive the second claim of (ii). If $|I|\log(\frac{en}{|I|}) \le \varrho|I_*|\log(\frac{en}{|I_*|}) - M$, then $|I\cup I_*|\log(\frac{en}{|I\cup I_*|}) \le |I|\log(\frac{en}{|I|}) + |I_*|\log(\frac{en}{|I_*|}) \le (1+\varrho)|I_*|\log(\frac{en}{|I_*|}) - M$. Hence, $|I_*|\log(\frac{en}{|I_*|}) \ge \frac{1}{1+\varrho}|I\cup I_*|\log(\frac{en}{|I\cup I_*|}) + \frac{M}{1+\varrho}$, which, together with the definition of the τ-oracle, imply

$$\begin{aligned}
\sum_{i\in I_*\setminus I}\frac{\theta_i^2}{\sigma^2} &\ge \Big(\sum_{i\in I^c}\frac{\theta_i^2}{\sigma^2} - \sum_{i\in I_*^c}\frac{\theta_i^2}{\sigma^2}\Big) \ge \tau_0\big(|I_*|\log(\tfrac{en}{|I_*|}) - |I|\log(\tfrac{en}{|I|})\big)\\
&\ge \tau_0(1-\varrho)|I_*|\log(\tfrac{en}{|I_*|}) + \tau_0 M \ge \tau'|I\cup I_*|\log(\tfrac{en}{|I\cup I_*|}) + \tfrac{2\tau_0}{1+\varrho}M,
\end{aligned} \tag{2.40}$$

where $\tau' = \frac{1-\varrho}{1+\varrho}\tau_0 > \bar\tau$ by the condition of the theorem. Thus, we obtain

$$\mathbb{E}_\theta\hat\pi\big(I : |I|\log(\tfrac{en}{|I|}) \le \varrho|I_*|\log(\tfrac{en}{|I_*|}) - M\big|X\big) \le \mathbb{E}_\theta\hat\pi\Big(I : \sum_{i\in I_*\setminus I}\frac{\theta_i^2}{\sigma^2} \ge \tau'|I\cup I_*|\log(\tfrac{en}{|I\cup I_*|}) + \tfrac{2\tau_0}{1+\varrho}M|X\Big).$$

By this and (2.39) with $I' = I_*$, the second claim of (ii) follows with $\alpha_1' = \frac{\alpha(\tau'-\bar\tau)}{3} > 0$ and $m_1' = \frac{2\tau_0 m_0'}{1+\varrho}$.

Finally, let us prove (iii). Denote $\mathcal{G}_3 = \mathcal{G}_3(\theta,M) = \{I : r^2(I,\theta) \ge c_3 r^2(\theta) + M\sigma^2\}$, where the constants $c_1 > 2$, c_2, c_3 are defined in Lemma 2.2. Applying Lemma 2.2 and using the fact (2.34), we complete the proof of (iii):

$$\mathbb{E}_\theta\hat\pi\big(I\in\mathcal{G}_3\big|X\big) = \sum_{I\in\mathcal{G}_3}\mathbb{E}_\theta\hat\pi(I|X) \le e^{-c_2M}\sum_{I\in\mathcal{I}}\big(\tfrac{ne}{|I|}\big)^{-c_1|I|} \le C_0 e^{-c_2M}. \qquad \square$$

Proof of Theorem 2.3. The biggest part of the proof is already contained in Theorem 2.2. We first establish the coverage property. The constants M_1, H_1 and m_1 are defined in Theorem 2.1, the constant ϱ is from (2.21). Take some $M_2 > \frac{M_1}{\varrho}$, for example $M_2 = \frac{M_1}{\varrho} + 1$. From (2.17) and (2.26), it follows that $r^2(\theta) \le r^2(I_*,\theta) = (b(\theta)+1)\sigma^2|I_*|\log(\frac{en}{|I_*|}) + b(\theta)\sigma^2 \le (b(\theta)+1)\sigma^2(|I_*|\log(\frac{en}{|I_*|}) + 1)$. Combining this with claims (ii) from Theorems 2.1 and 2.2

and the definition (2.24) of $\hat{r}$ yields the coverage property:

$$\begin{aligned}
&\mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, [(b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2]^{1/2}\big) \\
&\le \mathbb{P}_\theta\Big(\|\hat{\theta}-\theta\|^2 > (b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2, \hat{r}^2 \ge \varrho\sigma^2|I_*|\log(\tfrac{en}{|I_*|}) + \sigma^2 - \tfrac{M\sigma^2}{M_2}\Big) \\
&\quad + \mathbb{P}_\theta\Big(\hat{r}^2 < \varrho\sigma^2|I_*|\log(\tfrac{en}{|I_*|}) + \sigma^2 - \tfrac{M\sigma^2}{M_2}\Big) \\
&\le \mathbb{P}_\theta\Big(\|\hat{\theta}-\theta\|^2 > \varrho M_2 r^2(\theta) + M\sigma^2\Big) + \mathbb{P}_\theta\Big(|\hat{I}|\log(\tfrac{en}{|\hat{I}|}) < \varrho|I_*|\log(\tfrac{en}{|I_*|}) - \tfrac{M}{M_2}\Big) \\
&\le H_1\big(\tfrac{en}{|I_o|}\big)^{-\alpha_1|I_o|}e^{-m_1 M} + H_1'\big(\tfrac{en}{|I_*|}\big)^{-\alpha_1'|I_*|}e^{-m_1'' M} \le H_2 e^{-m_2 M},
\end{aligned}$$

where $\alpha_1 = \varrho M_2 - M_1$, $m_1'' = m_1'/M_2$, $H_2 = H_1 + H_1'$, $m_2 = m_1 \wedge m_1''$; H_1', α_1', m_1' are defined in Theorem 2.2 and the constant ϱ is from (2.21). The first claim of the theorem follows.

The size property follows from the definition (2.24) of $\hat{r}$, Remark 2.7 and property (i) of Theorem 2.2. Indeed, $\mathbb{P}_\theta\big(\hat{r}^2 \ge \sigma^2(M_0'+\gamma)|I_o|\log(\frac{en}{|I_o|}) + (M+1)\sigma^2\big) = \mathbb{P}_\theta\big(|\hat{I}|\log(\frac{en}{|\hat{I}|}) \ge (M_0'+\gamma)|I_o|\log(\frac{en}{|I_o|}) + M\big) \le \mathbb{P}_\theta\big(|\hat{I}|\log(\frac{en}{|\hat{I}|}) \ge M_0'|I\cap I_o|\log(\frac{en}{|I\cap I_o|}) + \gamma|I_o|\log(\frac{en}{|I_o|}) + M\big) \le H_0'(\frac{ne}{|I_o|})^{-\gamma|I_o|}e^{-M}$. □

Proof of Theorem 2.5. Observe that $r^2(\theta) \le r^2(I^*(\theta),\theta) \le \sigma^2 s(\theta)\log(\frac{en}{s(\theta)})$. Since the function $x \mapsto x\log(en/x)$ is increasing over $(0,n]$, $|I| \ge Ms(\theta)$ implies that

$$r^2(I,\theta) \ge \sigma^2|I|\log(\tfrac{en}{|I|}) \ge \sigma^2 Ms(\theta)\log(\tfrac{en}{Ms(\theta)}).$$

Thus, if $|I| \ge Ms(\theta)$, then

$$r^2(I,\theta) \ge M\sigma^2 s(\theta)\log(\tfrac{en}{Ms(\theta)}) \ge M_4 r^2(\theta) - M_4\sigma^2 s(\theta)\log(\tfrac{en}{s(\theta)}) + M\sigma^2 s(\theta)\log(\tfrac{en}{Ms(\theta)}).$$

The first claim follows from Theorem 2.2 with $M_4 = c_3$ and $m_4 = c_2$.

To prove the second claim, note that for any $M' > 2M_4$, $|I| \ge M's(\theta)$ implies that

$$r^2(I,\theta) \ge \sigma^2|I|\log(en/|I|) \ge M'\sigma^2 s(\theta)[\log(en/s(\theta)) - \log M'] \ge \frac{M'}{2}\sigma^2 s(\theta)\log(en/s(\theta)),$$

provided that $s(\theta) < en/(M')^2$. Since $r^2(\theta) \le r^2(I^*(\theta),\theta) \le \sigma^2 s(\theta)\log(en/s(\theta))$, the relation above implies that $r^2(I,\theta) \ge M_4 r^2(\theta) + M\sigma^2$, where $M = (M'/2 - M_4)s(\theta)\log(en/s(\theta))$. Hence by Theorem 2.2, the assertion holds for $M_4' = M'$ whenever $s(\theta) < en/(M')^2$. If $s(\theta) \ge en/(M')^2$, the result trivially holds by choosing $M_4' = (M')^2/e$. Hence the choice $M_4' \ge \max\{M', (M')^2/e\}$ ensures the result with $m_4' = m_4(M'/2 - M_4)$ for any $\theta \in \mathbb{R}^n$. □

Proof of Theorem 2.6. Recall (2.29): $\sigma^{-2}r^2(\theta) \le Kn(\frac{p_n}{n\sigma})^q\big[\log(\frac{n\sigma}{p_n})\big]^{1-q/2}$ for each $\theta \in m_q[p_n]$ with some $K = K(q)$. On the other hand, if $|I| > Mp_n^* = Men(\frac{p_n}{n\sigma})^q\big[\log(\frac{n\sigma}{p_n})\big]^{-q/2}$, then

$$\begin{aligned}
\sigma^{-2}r^2(I,\theta) &\ge |I|\log(\tfrac{en}{|I|}) \ge Mp_n^*\log(\tfrac{en}{Mp_n^*}) \\
&= Mp_n^*\big[q\log(\tfrac{n\sigma}{p_n}) + \frac{q}{2}\log\log(\tfrac{n\sigma}{p_n}) - \log(M)\big] \ge Mqp_n^*\log(\tfrac{n\sigma}{p_n})
\end{aligned}$$

for sufficiently large n as $p_n = o(n)$. Then, for any $\theta \in m_q[p_n]$, $M > c_3K/(qe)$ and $|I| > Mp_n^*$, we have that, for sufficiently large n,

$$\begin{aligned}\sigma^{-2}\big(r^2(I,\theta) - c_3 r^2(\theta)\big) &\geq Mp_n^* \log\big(\tfrac{en}{Mp_n^*}\big) - c_3 Kn\big(\tfrac{p_n}{n\sigma}\big)^q \big[\log\big(\tfrac{n\sigma}{p_n}\big)\big]^{1-q/2} \\ &\geq Mqp_n^* \log\big(\tfrac{n\sigma}{p_n}\big) - c_3 Ke^{-1} p_n^* \log\big(\tfrac{n\sigma}{p_n}\big) = (Mq - c_3 Ke^{-1}) p_n^* \log\big(\tfrac{n\sigma}{p_n}\big).\end{aligned}$$

Finally, applying Theorem 2.2, we obtain

$$\sup_{\theta \in m_q[p_n]} \mathbb{E}_\theta \hat{\pi}(I : |I| > Mp_n^* | X) \leq C_0 \exp\big\{-c_2 q\big(M - c_3 K(qe)^{-1}\big) p_n^* \log\big(\tfrac{n\sigma}{p_n}\big)\big\},$$

which gives the claim with $m_5 = c_2 q$ and $M_5 = c_3 K(qe)^{-1}$. □

3

LOCAL POSTERIOR CONCENTRATION RATE FOR MULTILEVEL SPARSE SEQUENCES

Suppose we observe $X = X^{(n)} = (X_1, \dots, X_n) \in \mathbb{R}^n$, with

$$X_i = \theta_i + \xi_i, \quad i \in \mathbb{N}_n = \{1, \dots, n\}, \tag{3.1}$$

where $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ is an unknown high-dimensional parameter of interest, the noise variables ξ_i's are independent standard Gaussian. In what follows, we let $n \geq 3$. The general goal is to make inference about θ based on the observed data X by using a Bayesian approach: in particular, recovery of the parameter θ and the derivation of the local contraction rate of the empirical Bayes posterior. We consider mainly the non-asymptotic results, which imply asymptotic assertions if needed.

In this canonical high-dimensional problem, useful inference is clearly not possible without some structure on the parameter. One of the popular structural assumptions is *sparsity*. In this chapter, we are concerned with a more generalized version of sparsity, namely, *multilevel sparsity*. The vector $\theta = (\theta_1, \dots, \theta_n)$ is assumed to be a *multilevel sparse*, i.e., the large proportion of the entries of θ consist of some values $a_1, \dots, a_m$. These values are known, but the proportions and the entries of θ at which these are taken are unknown. If $m = 1$ and $a_m = 0$, we obtain the traditional sparse signal.

One can extend the traditional sparsity class of *nearly black* vectors to multilevel sparsity class, but, to the best of our knowledge, multilevel sparsity structure is not considered in the literature and the minimax rate for this structure is not studied. For the traditional one-level ($m = 1$) sparsity structure, there is a variety of estimation methods and results are available in the literature: Donoho and Johnstone [38], Birgé and Massart [23], Johnstone and Silverman[50], Abramovich, Benjamini, Donoho and Johnstone [1], Abramovich, Grinshtein and Pensky [2], Castillo and van der Vaart [33], van der Pas, Kleijn and van der Vaart [85].

In this chapter, for inference on θ we use an empirical Bayes approach. Since any Bayesian approach always delivers a posterior $\pi(\vartheta|X)$ (in the posteriors for θ, we will use the variable ϑ to distinguish it from the "true" θ), an accompanying problem of interest is the contraction of the resulting (empirical Bayes) posterior to the "true" θ from the frequentist perspective of the "true" measure $\mathbb{P}_\theta$, the distribution of X from (3.1). The quality of posterior is characterized by the posterior contraction rate. We pursue the novel local approach, namely, the posterior contraction (and estimation) rate $r^2(\theta)$ is allowed to be a function of θ, i.e., it is a local quantity. The local approach is more flexible than the global one; more on this is in Section 3.3. The point is that we do not need to impose any specific sparsity structure on θ, because the proposed local approach automatically exploits the "effective" sparsity of each underlying θ. For instance, if θ happens to lie in a sparsity class (say, $\ell_0[p_n]$ or $m_s[p_n]$, see Section 3.3) and the sparsity level p_n is of polynomial order, then the adaptive (global) minimax results (in fact, for the two problems: estimation and posterior contraction rate) over the sparsity class follow from the local results. In particular, our local results imply the same type of certain (global) minimax estimation results over sparsity classes as in Johnstone and Silverman [50], and the same type of global minimax (over sparsity classes) results on contraction posterior rates as in Castillo and van der Vaart [33].

This chapter is organized as follows. In Section 3.1 we introduce the notation and describe the empirical Bayes procedure for multilevel sparse sequences. Section 3.2 contains the main results: the local (oracle) posterior contraction and estimation results for the constructed empirical Bayes posterior and the corresponding empirical Bayes posterior mean estimator, respectively, in terms of the local rate $r^2(\theta)$ uniformly over $\theta \in \mathbb{R}^n$. If the sparsity level is of polynomial order, then the global adaptive minimax (over sparsity classes) results on contraction posterior and estimation rates follow as a consequence of our local results. The implications of the local results, simulation study and proofs are collected in Sections 3.3, 3.4 and 3.5, respectively.

3.1. PRELIMINARIES

First we introduce some notation, then construct an empirical Bayes posterior.

3.1.1. NOTATION

Denote the probability measure of X from the model (3.1) by $\mathbb{P}_\theta = \mathbb{P}_\theta^{(n)}$. For the notational simplicity we often skip the dependence on n of this quantity and many others. Denote by I_{a_k} the index set of coordinates with a value a_k, by I the set of index coordinates with values which are not equal to $a_1, \dots, a_m$, so that $\mathrm{I} = (I^{a_1}, \dots, I^{a_m}, I)$ forms the partition of $\mathbb{N}_n = \{1, \dots, n\}$. Without loss of generality, we assume that $a_1 = 0$. Let M_n^m be the family of all possible partitions I, except for the partitions with $I = \varnothing$. Then $|\mathrm{M}_n^m| = (m+1)^n - m^n$.

3.1.2. EMPIRICAL BAYES POSTERIOR

As we mentioned before, we deal with the classical high-dimensional normal model $X = (X_i, i \in \mathbb{N}_n) \sim \mathbb{P}_\theta = \bigotimes_{i=1}^n \mathrm{N}(\theta_i, 1)$, $\theta = (\theta_i, i \in \mathbb{N}_n) \in \mathbb{R}^n$. We would like to design a prior that models multilevel sparse sequences θ with m levels. Namely, there are $m+1$

groups in vector $\theta = (\theta_{I^{a_1}}, \ldots, \theta_{I^{a_m}}, \theta_I)$, where $\theta_{I^{a_1}} = (\theta_i = a_1, i \in I^{a_1}), \ldots, \theta_{I^{a_m}} = (\theta_i = a_m, i \in I^{a_m}), \theta_I = (\theta_i, i \in I)$. It is reasonable to impose a prior on θ given the partition $\mathrm{I} = (I^{a_1}, \ldots, I^{a_m}, I)$ as follows:

$$\pi_{\mathrm{I}} = \bigotimes_{i=1}^{n} \mathrm{N}(\mu_i(\mathrm{I}), \tau_i^2(\mathrm{I})) = \Big[\bigotimes_{i \in I^{a_1}} \delta_{a_1}\Big] \times \ldots \times \Big[\bigotimes_{i \in I^{a_m}} \delta_{a_m}\Big] \times \Big[\bigotimes_{i \in I} \mathrm{N}(\mu_{m+1,i}, K)\Big], \tag{3.2}$$

where $\mu_i(\mathrm{I}) = \sum_{j=1}^{m} a_j \mathbb{1}\{i \in I^{a_j}\} + \mu_{m+1,i} \mathbb{1}\{i \in I\}$ and $\tau_i^2(\mathrm{I}) = K \mathbb{1}\{i \in I\}$, for some fixed $K > 0$. Next, we introduce the prior λ on $\mathrm{I} \in \mathrm{M}_n^m$ as follows: for $\kappa \geq 1$,

$$\lambda(\mathcal{I} = \mathrm{I}) = \lambda_{\mathrm{I}} = c_n \exp\big\{-\kappa[|I| + \sum_{j=2}^{m} |I^{a_j}|] \log n\big\}, \quad \mathrm{I} \in \mathrm{M}_n^m. \tag{3.3}$$

Since $\sum_{\mathrm{I} \in \mathrm{M}_n^m} \lambda_{\mathrm{I}} = 1$ and $|I| > 0$, the normalizing constant is $c_n = c_n(\kappa) = 1/\big[(1 + n^{-\kappa} + (m-1)n^{-\kappa}\mathbb{1}\{m \geq 2\})^n - (1 + (m-1)n^{-\kappa}\mathbb{1}\{m \geq 2\})^n\big]$. Putting a prior λ on M_n^m yields the resulting mixture prior for θ:

$$\pi = \sum_{\mathrm{I} \in \mathrm{M}_n^m} \lambda_{\mathrm{I}} \pi_{\mathrm{I}}, \tag{3.4}$$

where π_{I} is defined by (3.2). This in turn leads to the marginal distribution of X

$$\mathbb{P}_X = \sum_{\mathrm{I} \in \mathrm{M}_n^m} \lambda_{\mathrm{I}} \mathbb{P}_{X,\mathrm{I}}, \quad \mathbb{P}_{X,\mathrm{I}} = \bigotimes_{i=1}^{n} \mathrm{N}(\mu_i(\mathrm{I}), \tau_i^2(\mathrm{I}) + 1). \tag{3.5}$$

It remains to choose the parameters $\mu_{m+1,i}$ in the prior and we do this by using an empirical Bayes approach. The marginal likelihood $\mathbb{P}_X$ is readily maximized with respect to $\mu_{m+1,i}$: $\hat{\mu}_{m+1,i} = X_i$. Then we obtain the empirical Bayes posterior

$$\hat{\pi}(\vartheta|X) = \sum_{\mathrm{I} \in \mathrm{M}_n^m} \hat{\pi}(\vartheta, \mathcal{I} = \mathrm{I}|X) = \sum_{\mathrm{I} \in \mathrm{M}_n^m} \hat{\pi}(\vartheta|X, \mathcal{I} = \mathrm{I}) \hat{\pi}(\mathcal{I} = \mathrm{I}|X), \tag{3.6}$$

where the empirical Bayes conditional posterior (recall that $\mathrm{N}(0,0) = \delta_0$)

$$\hat{\pi}(\vartheta|X, \mathcal{I} = \mathrm{I}) = \hat{\pi}_{\mathrm{I}}(\vartheta|X) = \Big[\bigotimes_{i \in I^{a_1}} \delta_{a_1}\Big] \times \ldots \times \Big[\bigotimes_{i \in I^{a_m}} \delta_{a_m}\Big] \times \Big[\bigotimes_{i \in I} \mathrm{N}\Big(X_i, \tfrac{K}{K+1}\Big)\Big] \tag{3.7}$$

and the empirical Bayes posterior for $\mathrm{I} \in \mathrm{M}_n^m$

$$\hat{\pi}(\mathcal{I} = \mathrm{I}|X) = \frac{\lambda_{\mathrm{I}} \bigotimes_{i=1}^{n} \phi(X_i, \sum_{j=1}^{m} a_j \mathbb{1}\{i \in I^{a_j}\} + X_i \mathbb{1}\{i \in I\}, \tau_i^2(\mathrm{I}) + 1)}{\sum_{\mathrm{J} \in \mathrm{M}_n^m} \lambda_{\mathrm{J}} \bigotimes_{i=1}^{n} \phi(X_i, \sum_{j=1}^{m} a_j \mathbb{1}\{i \in J^{a_j}\} + X_i \mathbb{1}\{i \in J\}, \tau_i^2(\mathrm{J}) + 1)}. \tag{3.8}$$

Denoting $\hat{\mu}(\mathrm{I}) = \sum_{j=1}^{m} a_j \mathbb{1}\{i \in I^{a_j}\} + X_i \mathbb{1}\{i \in I\}, i \in \mathbb{N}_n$, we get an estimator based on $\hat{\pi}(\cdot|X)$, namely,

$$\hat{\theta} = \hat{\theta}(\mathrm{I}) = \hat{\mathbb{E}}(\vartheta|X) = \sum_{\mathrm{I} \in \mathrm{M}_n^m} \hat{\mu}(\mathrm{I}) \hat{\pi}(\mathcal{I} = \mathrm{I}|X), \tag{3.9}$$

which is nothing else but the *empirical Bayes mean*, with respect to the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ defined by (3.6).

3.2. MAIN RESULTS

In this section we introduce the local contraction rate for the empirical Bayes posterior $\hat{\pi}(\cdot|X)$. Notice that $\hat{\pi}(\cdot|X)$ is a random mixture over $\hat{\pi}_{\mathrm{I}}(\cdot|X), \mathrm{I} \in \mathrm{M}_n^m$. From the $\mathbb{P}_\theta$-perspective, each $\hat{\pi}_{\mathrm{I}}(\cdot|X)$ contracts to the true parameter θ with the local rate

$$\mathrm{R}^2(\mathrm{I},\theta) = \mathrm{R}^2(\mathrm{I},\theta,n,a_2,\dots,a_m) = \sum_{i\in I^{a_1}} \theta_i^2 + \sum_{j=2}^{m} \sum_{i\in I^{a_j}} (\theta_i - a_j)^2 + |I|, \quad \mathrm{I} \in \mathrm{M}_n^m. \tag{3.10}$$

Indeed, since $\hat{\mu}(\mathrm{I}) = \sum_{j=1}^m a_j \mathbb{1}\{i \in I^{a_j}\} + X_i \mathbb{1}\{i \in I\}, i \in \mathbb{N}_n$, the Markov inequality yields

$$\begin{aligned}\mathbb{E}_\theta \hat{\pi}_{\mathrm{I}}(\|\vartheta-\theta\|^2 \ge M^2\mathrm{R}^2(\mathrm{I},\theta)|X) &\le \frac{\mathbb{E}_\theta\big(\|\hat{\mu}(\mathrm{I})-\theta\|^2 + \frac{K|I|}{K+1}\big)}{M^2\mathrm{R}^2(\mathrm{I},\theta)} \\ &= \frac{(1+\frac{K}{K+1})|I| + \sum_{i\in I^{a_1}}\theta_i^2 + \sum_{j=2}^m \sum_{i\in I^{a_j}} (\theta_i - a_j)^2}{M^2\mathrm{R}^2(\mathrm{I},\theta)} \le \frac{2}{M^2}.\end{aligned}$$

For each $\theta \in \mathbb{R}^n$, there exists the best choice $\mathrm{I}_{\mathrm{o}} = \mathrm{I}_{\mathrm{o}}(\theta, a_2, \dots, a_m)$ of the partition $\mathrm{I} \in \mathrm{M}_n^m$ corresponding to the fastest local rate over the family of local rates $\mathrm{R}^2(\mathrm{M}_n^m) = \{\mathrm{R}^2(\mathrm{I},\theta), \mathrm{I} \in \mathrm{M}_n^m\}$: with $R^2(\mathrm{I},\theta)$ defined by (3.10),

$$\mathrm{R}^2(\theta) = \min_{\mathrm{I}\in\mathrm{M}_n^m} \mathrm{R}^2(\mathrm{I},\theta) = \sum_{i\in I_o^{a_1}} \theta_i^2 + \sum_{j=2}^{m} \sum_{i\in I_o^{a_j}} (\theta_i - a_j)^2 + |I_o|. \tag{3.11}$$

Ideally, we would like to have the quantity $\mathrm{R}^2(\theta)$ defined by (3.11) as the benchmark for the contraction rate of the empirical Bayes posterior $\hat{\pi}(\cdot|X)$ defined by (3.6). However, this turned out to be impossible, which is also confirmed by following estimation result of Donoho and Johnstone [37] for *sparse* signals:

$$\liminf_{n\to\infty} \frac{1}{\log n} \inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}_n} \Big[\frac{\mathbb{E}_\theta \|\theta - \hat{\theta}\|^2}{1 + \sum_{i=1}^n \min\{\theta_i^2, 1\}} \Big] \ge 2,$$

where the infimum is taken over all estimators, measurable functions of X. This shows that a reasonable benchmark for the contraction rate must contain a logarithmic factor, as was also shown by Birgé and Massart [23] for the estimation problem.

For a parameter $s > 1$, introduce the family of the so called *s-local rates* with a logarithmic factor:

$$r_m^2(\mathrm{I},\theta) = r_m^2(\mathrm{I},\theta,s,n,a_2,\dots,a_m) = \sum_{i\in I^{a_1}} \theta_i^2 + \sum_{j=2}^{m} \Big(\sum_{i\in I^{a_j}} (\theta_i - a_j)^2 + s^{-1}|I^{a_j}|\log n \Big) + |I|\log n,$$

where

$$a_k^2 \le K\log n, \quad k = 1,\dots,m, \quad for \quad some \quad K > 0. \tag{3.12}$$

There exists the best choice $\mathrm{I}_o = \mathrm{I}_o(\theta, s, n, a_2, \dots, a_m)$ of the partition $\mathrm{I} \in \mathrm{M}_n^m$ such that

$$\begin{aligned} r_m^2(\theta) &= r_m^2(\mathrm{I}_o,\theta) = \min_{\mathrm{I}\in\mathrm{M}_n^m} r_m^2(\mathrm{I},\theta) \\ &= \sum_{i\in I_o^{a_1}} \theta_i^2 + \sum_{j=2}^{m} \Big(\sum_{i\in I_o^{a_j}} (\theta_i - a_j)^2 + s^{-1}|I_o^{a_j}|\log n \Big) + |I_o|\log n. \end{aligned} \tag{3.13}$$

We call the quantity $r_m^2(\theta)$ and the choice I_o *oracle rate* and *oracle partition*, respectively. If $m = 1$, then the quantity $r_1^2(\theta)$ is nothing else but the oracle rate for *sparse* signals considered by Belitser and Nurushev [13]. It is easy to see that $r_m^2(\theta) \le r_1^2(\theta)$ for all $m > 1$. Indeed, this follows immediately, since $r_m^2(\theta) \le \min_{I\subseteq\mathbb{N}_n} r_m^2(\mathrm{I}_*(I),\theta) = \min_{\mathrm{I}\in\mathrm{M}_n^1} r_1^2(\mathrm{I},\theta)$, where $\mathrm{I}_*(I) = (I^c,\varnothing,\dots,\varnothing,I)$.

The following theorem establishes that the empirical Bayes posterior $\hat\pi(\vartheta|X)$ contracts to θ with the oracle rate $r_m^2(\theta)$ from the $\mathbb{P}_\theta$-perspective, and the empirical Bayes posterior mean $\hat\theta$, defined by (3.9), converges to θ with the oracle rate $r_m^2(\theta)$, uniformly over the entire parameter space.

Theorem 3.1. *Let the relation* (3.12) *be fulfilled. Then there exists a constant* $C_{or} = C_{or}(K,\kappa,m,s) > 0$ *such that for any* $\theta\in\mathbb{R}^n$ *and* $M > 0$,

$$\mathbb{E}_\theta\hat\pi\big(\|\vartheta-\theta\|^2 \ge M^2 r_m^2(\theta)|X\big) \le \frac{C_{or}}{M^2}, \tag{i}$$

$$\mathbb{E}_\theta\|\hat\theta-\theta\|^2 \le C_{or}\, r_m^2(\theta). \tag{ii}$$

The oracle interpretation of this result is as follows. A family of priors $\{\pi_{\mathrm{I}}, \mathrm{I}\in\mathrm{M}_n^m\}$ leads to the family of empirical Bayes posteriors $\{\hat\pi_{\mathrm{I}}(\vartheta|X), \mathrm{I}\in\mathrm{M}_n^m\}$. The best choice $\hat\pi_{\mathrm{I}_o}(\vartheta|X)$ (with the fastest (oracle) concentration rate $r_m^2(\theta)$) is not available to the observer, it can only be picked by the *oracle* who knows the true θ. We propose the mixture prior $\hat\pi(\vartheta|X)$ which does not use any knowledge of the oracle I_o. The above theorem says basically that the proposed empirical Bayes posterior $\hat\pi(\vartheta|X)$ *mimics the oracle* in the posterior contraction and estimation problems, i.e., $\hat\pi(\vartheta|X)$ performs as good as the oracle choice $\hat\pi_{\mathrm{I}_o}(\vartheta|X)$.

Remark 3.1. *Notice that we can make the oracle rate* (3.13) *smaller by choosing bigger values of the parameter s, but then the resulting constant* C_{or} *(which depends on s) will become bigger.*

3.3. IMPLICATIONS: THE MINIMAX RESULTS OVER SPARSITY CLASSES

We elucidate the potential strength of the oracle approach for sparse signals (i.e, $m = 1$). As we mentioned in Section 1.2, the local approach is more flexible than global in that local result imply a whole panorama of global minimax results for all sparsity scales (covered by the local rate) at once. Namely, suppose we have a sparsity scale $\{\Theta[p], p\in\mathcal{P}\}$ so that $\theta\in\Theta[p]$ with unknown sparsity parameter $p\in\mathcal{P}$. Next, suppose we established for some local rate $r(\theta)$ that

$$r(\theta) \le cR(\Theta[p]) \qquad \text{for all } \theta\in\Theta[p],\ p\in\mathcal{P}, \tag{3.14}$$

with some uniform $c > 0$. Then, clearly, the local results (for the posterior contraction and estimation problems) with local rate $r(\theta)$ will imply the global adaptive results *simultaneously for all scales* $\{\Theta[p], p\in\mathcal{P}\}$ with global rate $R(\Theta[p])$ for which (3.14) is satisfied. We say that the local rate $r(\theta)$ *covers* these scales.

Let us consider a couple of examples of sparsity scales $\{\Theta[p], p\in\mathcal{P}\}$ which are covered by our local rate $r_m(\theta) \le r_1(\theta)$ defined by (3.13). Let the conditions of Theorem 3.1 be fulfilled.

Nearly black vectors. Consider nearly black vectors with sparsity level $p_n = n^\gamma, \gamma \in (0,1)$ as $n \to \infty$,

$$\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \#(1 \le i \le n : \theta_i \ne 0) \le p_n\}.$$

It is a well-known fact that the minimax estimation rate over the class of nearly black vectors $\ell_0[p_n]$ is $R^2(\ell_0[p_n]) = 2p_n \log(n/p_n)(1+o(1))$ as $n \to \infty$ (see Donoho et al. [36]). For $p_n = n^\gamma$ with $\gamma \in (0,1)$, this reduces to $R^2(\ell_0[p_n]) = 2p_n \log(n/p_n)(1+o(1)) = O(p_n \log n)$.

We relate this minimax rate to the one-level oracle rate $r_1^2(\theta)$ (i.e., $m = 1$), $\theta \in \ell_0[p_n]$, by taking $\mathrm{I}_* = \mathrm{I}_*(\theta) = (I_*^c, I_*)$ with $I_* = I_*(\theta) = \{i \in \mathbb{N}_n : \theta_i \ne 0\}$:

$$\sup_{\theta \in \ell_0[p_n]} r_1^2(\theta) \le \sup_{\theta \in \ell_0[p_n]} r_1^2(\mathrm{I}_*, \theta) \le p_n \log n = O(R^2(\ell_0[p_n])).$$

We thus have the property (3.14) for $\Theta[p] = \ell_0[p_n]$. Hence, Theorem 3.1 immediately implies the adaptive minimax results on the estimation and contraction rate problems for the empirical Bayes posterior $\hat\pi(\vartheta|X)$. We summarize these results in the following corollary.

Corollary 3.1. *Let the empirical Bayes posterior $\hat\pi(\vartheta|X)$ be defined by (3.6) and $\hat\theta$ be defined by (3.9). Then there exist constants $C, c > 0$ (depending only on K, κ) such that for any $M > 0$,*

$$\sup_{\theta \in \ell_0[p_n]} \mathbb{E}_\theta \|\hat\theta - \theta\|^2 \le c\, p_n \log n,$$
$$\sup_{\theta \in \ell_0[p_n]} \mathbb{E}_\theta \hat\pi\big(\|\vartheta - \theta\|^2 \ge M p_n \log n | X\big) \le \frac{C}{M}.$$

Weak ℓ_s-balls. Consider weak ℓ_s-balls for $s \in (0,2)$ with sparsity level $p_n = n^\gamma, \gamma \in (0,1)$ as $n \to \infty$,

$$m_s[p_n] = \Big\{\theta \in \mathbb{R}^n : \tfrac{1}{n} \max_{1 \le i \le n} i|\theta_{[i]}|^s \le \big(\tfrac{p_n}{n}\big)^s\Big\},$$

where $|\theta_{[1]}| \ge \ldots \ge |\theta_{[n]}|$ are ordered values of $(|\theta_i|,\, i \in \mathbb{N}_n)$.

Denote $j = O_\theta(i)$ if $|\theta_i| = |\theta_{[j]}|$, with the convention that in the case $|\theta_{i_1}| = \ldots = |\theta_{i_k}|$ for $i_1 < \ldots < i_k$ we let $O_\theta(i_{l+1}) = O_\theta(i_l) + 1$, $l = 1, \ldots, k-1$. The minimax estimation rate over this class is $R^2(m_s[p_n]) = n(\frac{p_n}{n})^s \big(\log(\frac{n}{p_n})\big)^{(2-s)/2}(1+o(1))$ as $n \to \infty$ (see Donoho and Johnstone [38]). Since $p_n = n^\gamma, \gamma \in (0,1)$, $R^2(m_s[p_n]) = n(\frac{p_n}{n})^s \big(\log(\frac{n}{p_n})\big)^{(2-s)/2}(1+o(1)) = O(n^{1-s} p_n^s (\log n)^{(2-s)/2})$. Then, with $p_n^* = \big(\frac{p_n^2 n^{(2/s-2)}}{\log n}\big)^{s/2}$, $\mathrm{I}_* = \mathrm{I}_*(\theta) = (I_*^c, I_*)$, $I_* = I_*(\theta) = \{i \in \mathbb{N}_n : O_\theta(i) \le p_n^*\}$, we derive that for large enough n

$$\sup_{\theta \in m_s[p_n]} r_1^2(\theta) \le \sup_{\theta \in m_s[p_n]} r_1^2(\mathrm{I}_*, \theta) \le p_n^* \log n + p_n^2 n^{(2-2s)/s} \sum_{i=p_n^*+1}^{\infty} i^{-2/s}$$
$$\le p_n^* \log n + \tfrac{s}{2-s} p_n^2 n^{(2-2s)/s} p_n^{*(s-2)/s} \le c n^{1-s} p_n^s (\log n)^{(2-s)/2} = O(R^2(m_s[p_n])). \quad (3.15)$$

We established (3.14) for $\Theta[p] = m_s[p_n]$, thus Theorem 3.1 implies the next corollary.

Corollary 3.2. *Let the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ be defined by (3.6) and $\hat{\theta}$ be defined by (3.9). Then there exist constants $C, c > 0$ (depending only on K, κ) such that for any $M > 0$,*

$$\sup_{\theta \in m_s[p_n]} \mathbb{E}_\theta \hat{\pi}\big(\|\vartheta - \theta\|^2 \geq M n^{1-s} p_n^s (\log n)^{(2-s)/2} | X\big) \leq \frac{C}{M},$$
$$\sup_{\theta \in m_s[p_n]} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \leq c\, n^{1-s} p_n^s (\log n)^{(2-s)/2}.$$

Remark 3.2. *Recall that $r_m^2(\theta) \leq r_1^2(\theta) \leq R^2(\Theta)$ with both $\Theta = \ell_0[p_n]$ or $\Theta = m_s[p_n]$, for all $m \geq 2$, $a_2, \ldots, a_m$. Therefore by using multilevel sparsity model, we always improve upon the traditional minimax results for sparsity classes.*

3.4. SIMULATION STUDY

We simulated data according to the model (3.1) with dimension $n = 500$. We used signals $\theta = (\theta_1, \ldots, \theta_n)$ of the form $\theta = (a_1, \ldots, a_1, \ldots, a_m, \ldots, a_m, A, \ldots, A)$, where $a_1 = 0$ and the value A is assumed to be unknown. Denote the cardinality of a_j values in the signal θ by $N_{a_j}, j = 1, \ldots, m$. When performing simulations for the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ and some posterior based quantities, we used the values of the parameters $K = 10$ and $\kappa = 0.55$.

First, we did a small simulation study for the four estimators based on $\hat{\pi}(\vartheta|X)$: empirical Bayes posterior (EBP) mean given by (3.9) for multilevel sparse sequences ($a_1, a_2, \ldots, a_m$ are known values in advance, i.e., $m > 1$), EBP mean for one-level sparsity (only $a_1 = 0$ is known, $a_2, \ldots, a_m$ are unknown, i.e., $m = 1$) and the estimator $\check{\theta}$ (to be defined later) for multilevel sparse sequences and one-level sparsity, respectively. The construction of the estimator $\check{\theta}$ is straightforward (basically, it can be reduced to a hard thresholding estimator with a certain threshold). Computation of EBP mean, which is a shrinkage estimator, is a bit more involved. We provide some technical preliminaries for efficient computation of the mean and $\check{\theta}$ with respect to the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$.

EBP mean. According to (3.9), the EBP mean $\hat{\theta} = \int \vartheta d\hat{\pi}(\vartheta|X)$ is a random vector in $\mathbb{R}^n$. We can compute its ith coordinate as follows:

$$\hat{\theta}_i = \sum_{j=2}^{m} \frac{a_j \phi(X_i, a_j, 1)}{n^\kappa Q_i} + \frac{X_i}{n^\kappa \sqrt{2\pi(K+1)} Q_i}, \quad i = 1, \ldots, n,$$

where

$$Q_i = \phi(X_i, 0, 1) + \sum_{j=2}^{m} \frac{\phi(X_i, a_j, 1)}{n^\kappa} + \frac{1}{n^\kappa \sqrt{2\pi(K+1)}}, \quad i = 1, \ldots, n.$$

Estimator $\check{\theta}$. By applying the empirical Bayes approach with respect to I, we obtain that

$$\check{\mathrm{I}} = \operatorname*{argmax}_{\mathrm{I}\in\mathrm{M}_n^m} \hat{\pi}(\mathcal{I}=\mathrm{I}|X) = \operatorname*{argmax}_{\mathrm{I}\in\mathrm{M}_n^m} \Big\{ -\sum_{j=1}^{m}\sum_{i\in I^{a_j}} \frac{(X_i-a_j)^2}{2} + \log\lambda_{\mathrm{I}} - \frac{|I|}{2}\log(K+1)\Big\}$$
$$= \operatorname*{argmin}_{\mathrm{I}\in\mathrm{M}_n^m} \Big\{ \sum_{j=1}^{m}\sum_{i\in I^{a_j}} (X_i-a_j)^2 + 2\kappa\sum_{j=2}^{m}|I^{a_j}|\log n + |I|\big(2\kappa\log n + \log(K+1)\big)\Big\}. \tag{3.16}$$

Plugging in this into $\hat{\pi}_{\mathrm{I}}(\vartheta|X)$ defined by (3.7) gives the corresponding empirical (now "twice empirical": with respect to $\mu_{m+1,i}$ and with respect to I) Bayes estimator for θ:

$$\check{\theta} = \check{\theta}(\check{\mathrm{I}}) = \sum_{j=1}^{m} a_j \mathbb{1}\{i\in\check{I}^{a_j}\} + X_i\mathbb{1}\{i\in\check{I}\}, i\in\mathbb{N}_n. \tag{3.17}$$

Table 3.1 shows estimates of the mean square errors $\mathbb{E}_\theta\|\hat{\theta}-\theta\|^2$. These results are the average (square) error of 100 estimates $\hat{\theta}_1,\dots,\hat{\theta}_{100}$ computed from 100 data vectors simulated independently from the model (3.1). Besides, we also simulate the classical hard-thresholding *HT*, hard-thresholding oracle *HTO* and the empirical Bayes mean EBM considered by Johnstone and Silverman (2004) with a standard Laplace prior. The hard-thresholding *HT* and hard-thresholding oracle *HTO*, given by $\hat{\theta}_i^{HT} = X_i\mathbb{1}\{|X_i|>\sqrt{2\log n}\}$ and $\hat{\theta}_i^{HTO} = X_i\mathbb{1}\{|X_i|>\sqrt{2\log(n/p_n)}\}$. Note that the last estimator uses the "oracle" value of the sparsity parameter p_n, all the other estimators do not.

According to the results of Table 3.1, our estimators based on the *empirical Bayes posterior* $\hat{\pi}(\vartheta|X)$ are competitive to the other ones.

Table 3.1: *Average square errors of seven estimators computed on* 100 *data vectors X of length* $n=500$ *simulated from model* (3.1) *with* $\theta=(a_1,\dots,a_1,\dots,a_m,\dots,a_m,A,\dots,A)$.

Estmators	Average square errors $a_1=0,a_2=5,A=9$ $N_{a_1}=225,N_{a_2}=225,$ $N_A=50$	Average square errors $a_1=0,a_2=3,a_3=6,A=9$ $N_{a_1}=150,N_{a_2}=150,$ $N_{a_3}=150,N_A=50$
EBP mean, $m>1$	187	449
EBP mean, $m=1$	411	807
$\check{\theta}, m>1$	214	598
$\check{\theta}, m=1$	385	980
EBM	612	688
HT	607	1249
HTO	384	477

For further illustration in Figure 3.1 we visualize 95% credible intervals (gray bars) for θ with parameters $a_1=0,a_2=5,A=9,N_{a_1}=45,N_{a_2}=45,N_A=10,n=100$ and the empirical Bayes posterior means (red dots), by simulating 1000 draws from the empirical Bayes posterior distribution $\hat{\pi}(\vartheta|X)$ and plotting the 95% draws out of the 1000 that are closest to the EBP mean. Note that this picture shows good coverage.

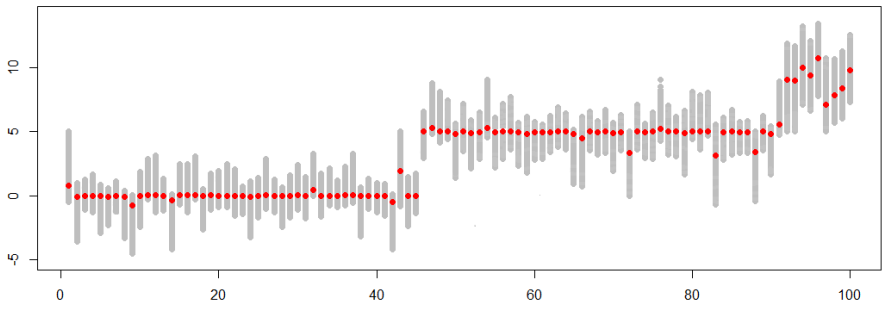

Figure 3.1: Empirical Bayes posterior means (red dots) and 95% credible intervals (gray bars) for the signal $\theta = (0,\dots,0,5,\dots,5,9,\dots,9)$ of length $n = 100$, where $N_0 = 45, N_5 = 45$ and $N_9 = 10$.

3.5. PROOFS

We provide a couple of technical lemmas used in the proof of the main result. For a $\tau_0 > 0$ and $\theta \in \mathbb{R}^n$, introduce the families of sets:

$$\mathcal{O}(\tau_0) = \mathcal{O}(\tau_0, \theta) = \{\mathrm{I} \in \mathrm{M}_n^m : r^2(\mathrm{I}, \theta) \le \tau_0 r^2(\theta)\}, \tag{3.18}$$

$$\mathcal{O}^c(\tau_0) = \mathcal{O}^c(\tau_0, \theta) = \{\mathrm{I} \in \mathrm{M}_n^m : r^2(\mathrm{I}, \theta) > \tau_0 r^2(\theta)\}, \tag{3.19}$$

where the oracle partition $\mathrm{I}_o = \mathrm{I}_o(\theta)$ is given by (3.13). The families $\mathcal{O}(\tau_0)$ and $\mathcal{O}^c(\tau_0)$ form a partition of M_n^m, as they do not intersect and $\mathrm{M}_n^m = \mathcal{O}(\tau_0) \cup \mathcal{O}^c(\tau_0)$. Denote for brevity $\hat{\pi}_\mathrm{I} = \hat{\pi}(\mathcal{I} = \mathrm{I}|X)$, where $\hat{\pi}(\mathcal{I} = \mathrm{I}|X)$ is defined by (3.8).

Lemma 3.1. *Let the measure $\hat{\pi}_\mathrm{I}$ be defined by (3.8), the oracle rate $r^2(\mathrm{I}_o, \theta)$ be defined by (3.13), $K > 0, s > 1$ and $\kappa > \max\{\frac{5}{9}\log 10, 0.9K\} + 2.49$. Then there exist positive constants $c_1 = c_1(\kappa) > 2, c_2$ and $c_3 = c_3(K, \kappa, s)$ such that*

$$\mathbb{E}_\theta \hat{\pi}_\mathrm{I} \le n^{-c_1(|I| + \sum_{j=2}^m |I^{a_j}|)} \exp\Big\{-c_2\big(r^2(\mathrm{I}, \theta) - c_3 r^2(\theta)\big)\Big\}.$$

Proof. Since $\hat{\pi}(X_1, \dots, X_n|\mathcal{I} = \mathrm{I})$ is the product of distributions $\mathrm{N}\big(\hat{\mu}_i(\mathrm{I}), \tau_i^2(\mathrm{I}) + 1\big)$, $i = 1, \dots, n$, where $\hat{\mu}_i(\mathrm{I}) = \sum_{j=1}^m a_j \mathbb{1}\{i \in I^{a_j}\} + X_i \mathbb{1}\{i \in I\}, i \in \mathbb{N}_n$ and $\tau_i^2(\mathrm{I}) = K\mathbb{1}\{i \in I\}$, with

densities $\phi\big(X_i, \hat{\mu}_i(\mathrm{I}), \tau_i^2(\mathrm{I})+1\big)$, we compute for any $h \in [0,1)$ and $\mathrm{I}_o \in \mathrm{M}_n^m$,

$$\begin{aligned}
\mathbb{E}_\theta \hat{\pi}_{\mathrm{I}} &= \mathbb{E}_\theta \frac{\lambda_{\mathrm{I}} \otimes_{i=1}^n \phi\big(X_i, \hat{\mu}_i(\mathrm{I}), \tau_i^2(\mathrm{I})+1\big)}{\sum_{\mathrm{J}\in \mathrm{M}_n^m} \lambda_{\mathrm{J}} \otimes_{i=1}^n \phi\big(X_i, \hat{\mu}_i(\mathrm{J}), \tau_i^2(\mathrm{J})+1\big)} \le \mathbb{E}_\theta \Big[\frac{\lambda_{\mathrm{I}} \otimes_{i=1}^n \phi\big(X_i, \hat{\mu}_i(\mathrm{I}), \tau_i^2(\mathrm{I})+1\big)}{\lambda_{\mathrm{I}_o} \otimes_{i=1}^n \phi\big(X_i, \hat{\mu}_i(\mathrm{I}_o), \tau_i^2(\mathrm{I}_o)+1\big)} \Big]^h \\
&= \mathbb{E}_\theta \Big[\frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{I}_o}} \Big]^h \exp\Big\{ -\sum_{j=1}^m \sum_{i \in I^{a_j}} \frac{h(X_i - a_j)^2}{2} + \sum_{k=1}^m \sum_{i \in I_o^{a_k}} \frac{h(X_i - a_k)^2}{2} + \frac{h(|I_o| - |I|)\log(K+1)}{2} \Big\} \\
&\le \mathbb{E}_\theta \Big[\frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{I}_o}} \Big]^h \exp\Big\{ -\sum_{j=1}^m \sum_{i \in I^{a_j} \setminus \cup_{k=1}^m I_o^{a_k}} \frac{h(X_i - a_j)^2}{2} + \sum_{k=1}^m \sum_{i \in I_o^{a_k} \setminus \cup_{j=1}^m I^{a_j}} \frac{h(X_i - a_k)^2}{2} \Big\} \\
&\quad \times \exp\Big\{ \sum_{j,k=1, j\neq k}^m \sum_{i \in I_o^{a_k} \cap I^{a_j}} \frac{h(2X_i a_j - 2X_i a_k + a_k^2 - a_j^2)}{2} + \frac{h|I_o|\log(K+1)}{2} \Big\}. \qquad (3.20)
\end{aligned}$$

Recall the elementary identity for $Y \sim \mathrm{N}(\mu, \sigma^2)$, $a, d \in \mathbb{R}$ and $b > -\sigma^2$:

$$\mathbb{E}[\exp\{-b(Y-a)^2/2\}] = \exp\Big\{ -\frac{(\mu - a)^2 b}{2(1+b\sigma^2)} - \frac{1}{2}\log(1+b\sigma^2) \Big\}, \qquad (3.21)$$

$$\mathbb{E}[\exp\{dY\}] = \exp\Big\{ \frac{d(d\sigma^2 + 2\mu)}{2} \Big\}. \qquad (3.22)$$

Now take $h = 0.9$ in (3.20). By using (3.21), we derive

$$\begin{aligned}
\mathbb{E}_\theta \hat{\pi}_{\mathrm{I}} \le & \Big[\frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{I}_o}} \Big]^{0.9} \exp\Big\{ -\frac{9}{38} \sum_{j=1}^m \sum_{i \in I^{a_j} \setminus \cup_{k=1}^m I_o^{a_k}} (\theta_i - a_j)^2 + 0.45|I_o|\log(K+1) \Big\} \\
& \times \exp\Big\{ 4.5 \sum_{k=1}^m \sum_{i \in I_o^{a_k} \setminus \cup_{j=1}^m I^{a_j}} (\theta_i - a_k)^2 + 0.5|I|\log 10 \Big\} \mathbb{E}_\theta e^{T(X)}, \qquad (3.23)
\end{aligned}$$

where $T(X) = 0.45 \sum_{j,k=1, j\neq k}^m \sum_{i \in I_o^{a_k} \cap I^{a_j}} (2X_i(a_j - a_k) + a_k^2 - a_j^2)$.

By using the relations (3.12) and (3.22), we obtain

$$\begin{aligned}
\mathbb{E}_\theta e^{T(X)} &= \mathbb{E}_\theta \exp\Big\{ 0.45 \sum_{j,k=1, j\neq k}^m \sum_{i \in I_o^{a_k} \cap I^{a_j}} \big(2X_i(a_j - a_k) + a_k^2 - a_j^2 \big) \Big\} \\
&= \exp\Big\{ 0.45 \sum_{j,k=1, j\neq k}^m \sum_{i \in I_o^{a_k} \cap I^{a_j}} \big((\theta_i - a_k)^2 - (\theta_i - a_j)^2 + 0.9(a_k - a_j)^2 \big) \Big\} \\
&\le \exp\Big\{ \sum_{j,k=1}^m \sum_{i \in I_o^{a_k} \cap I^{a_j}} \big(4.5(\theta_i - a_k)^2 - \frac{9}{38}(\theta_i - a_j)^2 \big) \Big\} \\
&\quad \times \exp\Big\{ 0.81K \big(\sum_{k=2}^m |I_o^{a_k}| + \sum_{j=2}^m |I^{a_j}| \big) \log n \Big\}. \qquad (3.24)
\end{aligned}$$

Denote the constants $c_0 = \max\{0.9K, \frac{5}{9}\log 10\} + \frac{5}{19} < \max\{0.9K, \frac{5}{9}\log 10\} + 2.49 < \kappa$ and

$c_1 = 0.9(\kappa - c_0) > 2$. The definition (3.3) of λ_{I}, $n \geq 3$ and $s > 1$ entail that

$$\Big[\frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{I}_o}}\Big]^{0.9} \exp\Big\{0.5|I|\log 10 + 0.81K\big(\sum_{j=2}^{m}|I^{a_j}|\big)\log n\Big\}$$
$$\leq \exp\Big\{-\big[(c_1+\tfrac{9}{38})|I| + (c_1+\tfrac{9}{38s})\big(\sum_{j=2}^{m}|I^{a_j}|\big)\big]\log n + 0.9\kappa(|I_o| + \tfrac{1}{s}\sum_{k=2}^{m}|I_o^{a_k}|)\log n\Big\}.$$

Using the relations (3.23), (3.24) and the last inequality, we derive that

$$\mathbb{E}_\theta\hat{\pi}_{\mathrm{I}} \leq n^{-c_1(|I|+\sum_{j=2}^{m}|I^{a_j}|)} \exp\Big\{-\tfrac{9}{38}\big(\sum_{j=1}^{m}\sum_{i\in I^{a_j}}(\theta_i - a_j)^2 + (|I| + \tfrac{1}{s}\sum_{j=2}^{m}|I^{a_j}|)\log n\big)\Big\}$$
$$\times \exp\Big\{C\big(\sum_{k=1}^{m}\sum_{i\in I_o^{a_k}}(\theta_i - a_k)^2 + (|I_o| + \tfrac{1}{s}\sum_{k=2}^{m}|I_o^{a_k}|)\log n\Big\},$$

where $C = C(K,\kappa,s) = s\max\{4.5, 0.9\kappa + 0.81K\}$. This completes the proof, with the constants $c_2 = \frac{9}{38}$ and $c_3 = c_3(K,\kappa,s) = \frac{38s}{9}\max\{4.5, 0.9\kappa + 0.81K\}$. □

Lemma 3.2. *Let $\theta \in \mathbb{R}^n$ and let $\hat{\theta}(\mathrm{I}), \mathrm{I} \in \mathrm{M}_n^m$, be defined by (3.9), and the set $\mathcal{O}(\tau_0)$ by (3.18). Then*

$$\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\|\hat{\theta}(\mathrm{I}) - \theta\|^2\hat{\pi}_{\mathrm{I}}\Big] \leq 6\tau_0 r^2(\theta).$$

Proof. Recall that $\xi_i = (X_i - \theta_i) \overset{ind}{\sim} \mathrm{N}(0,1)$, $i \in \mathbb{N}_n$, under $X \sim \mathbb{P}_\theta$. Write

$$\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\|\hat{\theta}(I) - \theta\|^2\hat{\pi}_{\mathrm{I}}\Big] = \mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\Big(\sum_{i\in I}\xi_i^2 + \sum_{j=1}^{m}\sum_{i\in I^{a_j}}(\theta_i - a_j)^2\Big)\hat{\pi}_{\mathrm{I}}\Big]$$
$$\leq \mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\Big(\sum_{i\in I}\xi_i^2\Big)\hat{\pi}_{\mathrm{I}}\Big] + \tau_0 r^2(\theta). \tag{3.25}$$

It is known fact that

$$\exp\{t\mathbb{E}[\max_{1\leq i\leq n}\xi_i^2]\} \leq \mathbb{E}\exp\{t\max_{1\leq i\leq n}\xi_i^2\} \leq \sum_{i=1}^{n}\mathbb{E}\exp\{t\xi_i^2\} = \frac{n}{\sqrt{1-2t}}.$$

Here we used Jensen's inequality. Therefore $\mathbb{E}[\max_{1\leq i\leq n}\xi_i^2] \leq \frac{\log n}{t} - \frac{\log(1-2t)}{2t}$. Taking $t = \frac{2}{5}$, we derive that for any $n \geq 3$

$$\mathbb{E}[\max_{1\leq i\leq n}\xi_i^2] \leq \frac{5\log n}{2} + \frac{5\log 5}{4} \leq \Big(\frac{5}{2} + \frac{5\log 5}{4\log 3}\Big)\log n < 5\log n. \tag{3.26}$$

Since $\mathrm{I} \in \mathcal{O}(\tau_0)$, it is not difficult to see that $|I| \leq \frac{\tau_0 r^2(\theta)}{\log n}$. Applying this and (3.26), we obtain

$$\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\Big(\sum_{i\in I}\xi_i^2\Big)\hat{\pi}_{\mathrm{I}}\Big] \leq \mathbb{E}_\theta\Big[\max_{1\leq i\leq n}\xi_i^2\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\Big(|I|\hat{\pi}_{\mathrm{I}}\Big)\Big]$$
$$\leq \frac{\tau_0 r^2(\theta)}{\log n}\mathbb{E}_\theta\Big[\max_{1\leq i\leq n}\xi_i^2\Big] \leq 5\tau_0 r^2(\theta). \tag{3.27}$$

Combining the last relation with (3.25) completes the proof of the lemma. □

Now we are ready to prove the main result, Theorem 3.1.

Proof of Theorem 3.1. Let $\hat{\mathbb{E}}$ and $\widehat{\mathrm{var}}$ denote the (random) expectation and variance with respect to $\hat{\pi}(\vartheta|X,\mathcal{I}=I)$ given by (3.7). Then from (3.7), it follows that

$$\hat{\mathbb{E}}\big(\|\vartheta-\theta\|^2|X,\mathcal{I}=I\big)=\sum_{i\in\mathbb{N}_n}\widehat{\mathrm{var}}(\vartheta_i|X,\mathcal{I}=I)+\sum_{i\in\mathbb{N}_n}\big(\hat{\mathbb{E}}(\vartheta_i|X,\mathcal{I}=I)-\theta_i\big)^2$$
$$=\frac{K|I|}{K+1}+\sum_{i\in I}\xi_i^2+\sum_{j=1}^{m}\sum_{i\in I^{a_j}}(\theta_i-a_j)^2\le r^2(I,\theta)+\sum_{i\in I}\xi_i^2,$$

where $\xi_i=(X_i-\theta_i)\sim\mathrm{N}(0,1)$.

The last relation and the Markov inequality imply that

$$\mathbb{E}_\theta\hat{\pi}\big(\|\vartheta-\theta\|\ge Mr(\theta)|X\big)=\mathbb{E}_\theta\sum_{\mathrm{I}\in\mathrm{M}_n^m}\hat{\pi}\big(\|\vartheta-\theta\|\ge Mr(\theta)|X,\mathcal{I}=\mathrm{I}\big)\hat{\pi}_\mathrm{I}$$
$$\le\mathbb{E}_\theta\sum_{\mathrm{I}\in\mathrm{M}_n^m}\frac{\hat{\mathbb{E}}\big(\|\vartheta-\theta\|^2|X,\mathcal{I}=\mathrm{I}\big)}{M^2r^2(\theta)}\hat{\pi}_\mathrm{I}\le\frac{\sum_{\mathrm{I}\in\mathrm{M}_n^m}r^2(\mathrm{I},\theta)\mathbb{E}_\theta\hat{\pi}_\mathrm{I}}{M^2r^2(\theta)}+\frac{\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathrm{M}_n^m}\big(\sum_{i\in I}\xi_i^2\big)\hat{\pi}_\mathrm{I}\Big]}{M^2r^2(\theta)}.\quad(3.28)$$

Let the sets $\mathcal{O}(\tau_0)$ and $\mathcal{O}^c(\tau_0)$ be defined by (3.18) and (3.19), respectively. Let τ_0 be chosen in such a way that $\tau_0>c_3=(38s/9)\max\{4.5,0.9\kappa+0.81K\}$ is defined in the proof of Lemma 3.1 and $\kappa>\max\big\{0.9K,\frac{5}{9}\log 10\big\}+2.49$. For $\mathrm{I}\in\mathcal{O}^c(\tau_0)$, we evaluate

$$r^2(\mathrm{I},\theta)-c_3r^2(\theta)\ge\big(1-\tfrac{c_3}{\tau_0}\big)r^2(\mathrm{I},\theta).\quad(3.29)$$

Denote $B=B(K,\kappa,s,\tau_0)=\frac{c_2(\tau_0-c_3)}{2\tau_0}=\frac{9(\tau_0-c_3)}{76\tau_0}$, where $c_2=\frac{9}{38}$ is defined in the proof of Lemma 3.1. Using Lemma 3.1, (3.29) and the facts that $\max_{x\ge0}\{xe^{-cx}\}\le(ce)^{-1}$ (for any $c>0$) and $(1+mn^{-c_1/2})^n\le e^m$, we obtain that

$$\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}r^2(\mathrm{I},\theta)\big[\mathbb{E}_\theta\hat{\pi}_I\big]^{\frac12}\le\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}r^2(\mathrm{I},\theta)n^{-c_1(|I|+\sum_{j=2}^m|I^{a_j}|)}\exp\Big\{-c_2\big(r^2(\mathrm{I},\theta)-c_3r^2(\theta)\big)\Big\}$$
$$\le\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}n^{-c_1(|I|+\sum_{j=2}^m|I^{a_j}|)}r^2(\mathrm{I},\theta)e^{-Br^2(\mathrm{I},\theta)}\le\frac{1}{Be}\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}n^{-c_1(|I|+\sum_{j=2}^m|I^{a_j}|)}$$
$$\le\frac{1}{Be}\sum_{k_1+k_2+\ldots+k_{m+1}=n}\binom{n}{k_1,k_2,\ldots,k_{m+1}}n^{-\frac{c_1\sum_{j=2}^{m+1}k_j}{2}}\le\frac{e^{m-1}}{B},\quad(3.30)$$

where $c_1=c_1(\kappa)=0.9(\kappa-c_0)>2$ is defined in the proof of Lemma 3.1.

If $\mathrm{I}\in\mathcal{O}(\tau_0)$, then $|I|\le\frac{\tau_0r^2(\theta)}{\log n}$. Combining this with Lemma 3.2 yields

$$\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\hat{\pi}_\mathrm{I}\sum_{i\in I}\xi_i^2\Big]\le\mathbb{E}_\theta\Big[\max_{1\le i\le n}\xi_i^2\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}|I|\hat{\pi}_\mathrm{I}\Big]\le\frac{\tau_0r^2(\theta)}{\log n}\mathbb{E}_\theta\Big[\max_{1\le i\le n}\xi_i^2\Big]\le5\tau_0r^2(\theta).\quad(3.31)$$

We have $\mathbb{E}(\sum_{i\in I}\xi_i^2)^2=|I|^2+2|I|\le3|I|^2$. Using this, Cauchy-Schwarz inequality and (3.30),

we evaluate

$$\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}\hat\pi_{\mathrm{I}}\sum_{i\in I}\xi_i^2\Big]\le\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}\Big[\mathbb{E}_\theta\Big(\sum_{i\in I}\xi_i^2\Big)^2\Big]^{\frac12}\Big[\mathbb{E}_\theta\hat\pi_{\mathrm{I}}^2\Big]^{\frac12}$$
$$\le\sqrt{3}\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}r^2(\mathrm{I},\theta)\Big[\mathbb{E}_\theta\hat\pi_{\mathrm{I}}\Big]^{\frac12}\le\frac{\sqrt{3}e^{m-1}}{B}. \tag{3.32}$$

From (3.31) and (3.32), it follows that

$$\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathrm{M}_n^m}\Big(\hat\pi_{\mathrm{I}}\sum_{i\in I}\xi_i^2\Big)\Big]=\mathbb{E}_\theta\Big[\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}\Big(\hat\pi_{\mathrm{I}}\sum_{i\in I}\xi_i^2\Big)+\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}\Big(\hat\pi_{\mathrm{I}}\sum_{i\in I}\xi_i^2\Big)\Big]$$
$$\le5\tau_0r^2(\theta)+\frac{\sqrt{3}e^{m-1}}{B}. \tag{3.33}$$

Recall that $\sum_{\mathrm{I}}\hat\pi_{\mathrm{I}}=1$ and $r^2(I,\theta)\le\tau_0r^2(\theta)$ for all $\mathrm{I}\in\mathcal{O}(\tau_0)$. Using these relations and (3.30), we have

$$\sum_{\mathrm{I}\in\mathrm{M}_n^m}r^2(\mathrm{I},\theta)\mathbb{E}_\theta\hat\pi_{\mathrm{I}}=\sum_{\mathrm{I}\in\mathcal{O}(\tau_0)}r^2(\mathrm{I},\theta)\mathbb{E}_\theta\hat\pi_{\mathrm{I}}+\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}r^2(\mathrm{I},\theta)\mathbb{E}_\theta\hat\pi_{\mathrm{I}}$$
$$\le\tau_0r^2(\theta)+\sum_{\mathrm{I}\in\mathcal{O}^c(\tau_0)}r^2(\mathrm{I},\theta)\mathbb{E}_\theta\hat\pi_{\mathrm{I}}\le\tau_0r^2(\theta)+\frac{e^{m-1}}{B}. \tag{3.34}$$

Finally, combining the relations (3.28), (3.33) and (3.34), and taking into account that $r^2(\theta)\ge1$, we finish the proof of assertion (i):

$$\mathbb{E}_\theta\hat\pi\big(\|\vartheta-\theta\|^2\ge M^2r^2(\theta)|X\big)\le\frac{6\tau_0}{M^2}+\frac{(\sqrt{3}+1)e^{m-1}}{M^2r^2(\theta)B}\le\frac{C_{or}}{M^2},$$

where $C_{or}=6\tau_0+\frac{(\sqrt{3}+1)e^{m-1}}{B}=6\tau_0+\frac{76\tau_0(\sqrt{3}+1)e^{m-1}}{9(\tau_0-c_3)}$, and we take, say, $\tau_0=c_3+1$.

The proof of assertion (ii) is essentially contained in the proof of the first assertion (i). Indeed, notice from (3.28), (3.33) and (3.34) that we proved a slightly stronger bound

$$\mathbb{E}_\theta\hat{\mathbb{E}}(\|\vartheta-\theta\|^2|X)=\mathbb{E}_\theta\sum_{\mathrm{I}\in\mathrm{M}_n^m}\hat{\mathbb{E}}\big(\|\vartheta-\theta\|^2|X,\mathcal{I}=\mathrm{I}\big)\hat\pi_{\mathrm{I}}\le C_{or}r^2(\theta).$$

This bound and $\|\hat\theta-\theta\|^2=\|\hat{\mathbb{E}}(\vartheta|X)-\theta\|^2\le\hat{\mathbb{E}}(\|\vartheta-\theta\|^2|X)$ imply the second assertion (ii): $\mathbb{E}_\theta\|\hat\theta-\theta\|^2\le\mathbb{E}_\theta\hat{\mathbb{E}}(\|\vartheta-\theta\|^2|X)\le C_{or}r^2(\theta)$. □

4

LOCAL INFERENCE BY PENALIZATION METHOD FOR BICLUSTERING MODEL

Suppose we observe a matrix $X = (X_{ij}) \in \mathbb{R}^{\boldsymbol{n}}$, $\boldsymbol{n} = (n_1, n_2) \in \mathbb{N}^2$:

$$X_{ij} = \theta_{ij} + \sigma\xi_{ij}, \quad i = 1,\dots,n_1, \quad j = 1,\dots,n_2, \tag{4.1}$$

where $\theta = (\theta_{ij}) \in \mathbb{R}^{\boldsymbol{n}}$ is an unknown high-dimensional parameter of interest with *biclustering* structure, $\xi = (\xi_{ij}) \in \mathbb{R}^{\boldsymbol{n}}$ is a random matrix with $\mathbb{E}_\theta \xi_{ij} = 0$. In general, ξ depends on θ, but we often suppress this dependence in the notation. Without loss of generality let $\sigma^2 = 1$. The general goal is to make inference about the parameter θ based on the data X: recovery of θ and *uncertainty quantification* by constructing an *optimal confidence set*. We pursue the *robust inference* in the sense that the distribution of ξ is unknown (can also depend on θ), the ξ_{ij}'s are in general not identically distributed and even not independent, but assumed to satisfy only certain mild condition (Condition (C1) in Section 4.1).

The inference problem is unfeasible unless some structure on θ is imposed. In this chapter we are concerned with *biclustering structure* which was first introduced in [47]. The essence of biclustering structure is to reduce dimensionality of a large matrix of parameters by simultaneous grouping of the rows and columns. For example, if the rows of θ correspond to objects and the columns to features, a biclustering structure means that only a few features are relevant for identifying a few groups of similar objects. Precisely, biclustering structure means that the rows and columns of the matrix $\theta = (\theta_{ij}) \in \mathbb{R}^{\boldsymbol{n}}$ are split into k_1 and k_2 clusters, respectively, and the values θ_{ij} are the same for i, j from the same clusters.

A particular case of biclustering structure is the *stochastic block model* (SBM) which is a popular model for the network analysis. In SBM $n_1 = n_2 = n$, row clusters coincide with the column clusters, indices i, j denote the vertices in an undirected net-

work graph of n vertices, clusters have the meaning of communities, and the observations $X_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$ for $i > j$, independent Bernoulli random variables with success probabilities $\theta_{ij} \in [0,1]$. This is the model (4.1) with $\xi_{ij} \in \{1-\theta_{ij}, -\theta_{ij}\}$ and $\mathbb{P}(\xi_{ij} = 1-\theta_{ij}) = \theta_{ij}$. Here X_{ij} stands for the presence or absence of an edge between vertices i and j. Typically, to model undirected network graphs, additional structure is imposed: $X_{ii} = \theta_{ii} = 0$, $X_{ij} = X_{ji}$ and $\theta_{ij} = \theta_{ji}$. Thus, we observe the adjacency matrix $X = (X_{ij})$ and want to infer on the symmetric block constant probability matrix $\theta = (\theta_{ij})$. The SBM is discussed in detail in Section 4.3.2.

The best studied problem is that of estimating θ in biclustering, stochastic block models and graphon classes (especially in the network analysis when the errors are independent Bernoulli): [3], [4], [68], [69]. Recently, minimax estimation results were derived in [40], [41] and [54]. However, even an optimal estimator does not reveal how far it is from θ. It is of great importance to quantify this uncertainty, which can be cast into the problem of constructing confidence sets for the parameter θ. We apply a penalization method with a suitably chosen penalty. The method delivers estimators for θ and the structural parameter (the partitions in row and column clusters), which we use for constructing a confidence ball.

For the usual norm $\|\cdot\|$ in $\mathbb{R}^{\boldsymbol{n}}$ (which is the Frobenius norm $\|\theta\|_F$ if θ is seen as $(n_1 \times n_2)$-matrix), a random ball in $\mathbb{R}^{\boldsymbol{n}}$ is $B(\hat{\theta}, \hat{r}) = \{\theta \in \mathbb{R}^{\boldsymbol{n}} : \|\hat{\theta} - \theta\| \le \hat{r}\}$, where the center $\hat{\theta} = \hat{\theta}(X) : \mathbb{R}^{\boldsymbol{n}} \mapsto \mathbb{R}^{\boldsymbol{n}}$ and radius $\hat{r} = \hat{r}(X) : \mathbb{R}^{\boldsymbol{n}} \mapsto \mathbb{R}_+ = [0, +\infty]$ are measurable functions of the data X. We consider non-asymptotic results which imply asymptotic assertions if needed. Denote by $N_{\boldsymbol{n}} = n_1 n_2$, the total number of the observations in the model (4.1) and the dimension of the space $\mathbb{R}^{\boldsymbol{n}}$. A traditional asymptotic regime is the high-dimensional setting $\dim(\mathbb{R}^{\boldsymbol{n}}) = n_1 n_2 = N_{\boldsymbol{n}} \to \infty$. One can also consider asymptotic settings for $n_1, n_2 \to \infty$ in different ways, or the asymptotic regime $\sigma \to 0$ (in case $\sigma^2 \ne 1$ in the model (4.1)); or various combinations of all the mentioned; e.g., $\sigma = \sigma_{\boldsymbol{n}} \to 0$ as $N_{\boldsymbol{n}} \to \infty$.

Besides estimating the unknown parameter, we wish to solve uncertainty quantification for biclustering model. The optimality framework for uncertainty quantification is already discussed in detail in Section 1.4. To the best of our knowledge, there are no results on uncertainty quantification (1.14) for biclustering structures. This chapter of thesis attempts to fill this gap and also to extend to the misspecified models as we allow the ξ_{ij}'s to be not necessarily independent Bernoulli (or normal). Moreover, we pursue the novel local approach, namely, the radial rate $r(\theta)$ in (1.14) is allowed to be a function of θ, which, in a way, measures the amount of biclustering structure in each $\theta \in \mathbb{R}^{\boldsymbol{n}}$: the smaller $r(\theta)$, the more structured θ is. The local approach is certainly more flexible than global, because we do not need to impose any specific biclustering structure and the proposed local approach automatically exploits the "effective" biclustering structure of each underlying θ (not necessarily having an exact biclustering structure).

Let us describe the main results of Chapter 4. First we propose a general penalization method for estimating the parameter of the biclustering model. For the proposed estimator $\hat{\theta}$, we derive the local estimation result in a refined non-asymptotic formulation: uniformly in $\theta \in \mathbb{R}^{\boldsymbol{n}}$, $\sup_{\theta \in \mathbb{R}^{\boldsymbol{n}}} \mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 \ge M_1 r^2(\theta) + M\big) \le H_1 e^{-m_1 M}$ for some fixed $M_1, H_1, m_1 > 0$ and arbitrary $M \ge 1$, an exponential non-asymptotic concentration bound in terms of M. This formulation provides a rather refined characterization of the

quality of the estimator $\hat{\theta}$ (finer than traditional oracle inequalities in expectation) and allows to study various asymptotic regimes. This oracle estimation result, besides being an ingredient for the uncertainty quantification problem (1.14), is of interest and importance on its own as it says in essence that the estimator $\hat{\theta}$ converges to θ with the local (oracle) rate $r(\theta)$.

Next, we construct a confidence ball by using the penalization method. Under some additional assumption, we obtain a rather strong type of optimality for this confidence ball. Namely, we establish the coverage and size relations in the optimality framework (1.14) with $\Theta_0 = \Theta_1 = \mathbb{R}^{\boldsymbol{n}}$ and the effective local radial rate $r(\theta) + N_{\boldsymbol{n}}^{1/4}$ (where $r(\theta)$ is the local oracle rate derived in the estimation result). Interestingly, this actually means that there is (almost) no deceptiveness issue for the biclustering model. Indeed, there is no payment in terms of removing deceptive parameters from the parameter set $\mathbb{R}^{\boldsymbol{n}}$ in the coverage relation, and the size relation holds also uniformly over $\Theta_1 = \mathbb{R}^{\boldsymbol{n}}$. Ideally, we would want the size relation to hold with the effective radial rate $r(\theta)$, but there is an extra term $N_{\boldsymbol{n}}^{1/4}$ in the expression for the effective radial rate $r(\theta) + N_{\boldsymbol{n}}^{1/4}$. This is not a problem for the "majority" of θ's as this extra term does not increase the order of the radial rate: $N_{\boldsymbol{n}}^{1/4} \le c\, r(\theta)$ for all $\theta \in \mathbb{R}^{\boldsymbol{n}} \setminus \tilde{\Theta}$ for some "thin" set $\tilde{\Theta}$. The set $\tilde{\Theta}$ can be informally described as a set of "highly structured" parameters: namely, when either the (oracle) number of row or column blocks of the underlying parameter θ is 1. We elaborate on this in Section 4.2.2.

We can conclude that, modulo this minor issue with the set $\tilde{\Theta}$ of highly structured parameters, the biclustering model in the present setting (almost) does not suffer from the deceptiveness problem. This is peculiar for the biclustering model and an intuitive explanation for this is that the biclustering model is too difficult (or too uninformative) to (fully) suffer from the deceptiveness problem. The additional assumption needed for the main result on uncertainty quantification is that we have a second sample of observations independent of the first sample (the observations can be dependent within samples) with the error ξ satisfying some mild natural conditions. In the case of independent normal errors in the model (4.1), this assumption is automatically fulfilled. Indeed, in this case we can "duplicate" the observations by randomization at the cost of doubling the variance in the following manner: create samples $X' = X + Z$ and $X'' = X - Z$, for a $Z = (Z_{ij}, (i,j) \in [\boldsymbol{n}])$ (independent of X) such that $Z_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. Admittedly, the assumption about having a second sample is a bit awkward, but we were unable to establish the result on uncertainty quantification in the general case without this assumption. It is however the question whether this assumption is really essential for the general case or it is just an artifact of our proof technique.

An important consequence of our local approach is that a panorama of adaptive (global) minimax results (for the both estimation and uncertainty quantification problems) over *all* biclustering and graphon classes *covered* by $r(\theta)$ (see Section 4.3) follows from our local results. In particular, our results imply the same type of adaptive minimax estimation results for the biclustering model as in [41]. The relation of our estimation results to the results of [41] is discussed in Remark 4.14. The results from [40] and [54] for the stochastic block model (with implications for network modeling) and graphon classes follow as well since the stochastic block model is a particular case of the biclustering model.

The Chapter 4 is organized as follows. In Section 4.1 we introduce the notation and describe the proposed penalization procedure in detail. Section 4.2 contains the main results of Chapter 4. Section 4.3 contains the implications of our local results for adaptive minimax results on the both problems: estimation and uncertainty quantification problems. The proofs of the lemmas and theorems are gathered in Sections 4.4 and 4.5.

4.1. PRELIMINARIES

In this section we first introduce some notation. Next, we start with a heuristic motivation of the key technical condition, then provide the condition itself. The section is completed with the *penalization method* which we use in the construction of the estimator and the confidence ball.

At first reading, one may want to skip this section (only referring to some definitions from this section) and go ahead to Section 4.2 which contains the main results of Chapter 4.

4.1.1. NOTATION

For $m_1, m_2 \in \mathbb{N}$, denote $[m_1] = \{1, 2, \dots, m_1\}$, $\boldsymbol{m} = (m_1, m_2)$, $[\boldsymbol{m}] = ([m_1], [m_2])$. For an $(m_1 \times m_2)$-matrix $x = (x_{ij}) \in \mathbb{R}^{\boldsymbol{m}}$, we will interchangeably use the same notation x to denote the vector $x = \mathrm{vec}\big[(x_{ij})\big] = (x_{11}, x_{12} \dots, x_{n_1 n_2})^T$. Conversely, for any $x \in \mathbb{R}^{m_1 m_2}$ we can use matricized indexing $x = (x_{11}, x_{12}, \dots, x_{m_1 m_2})^T$. Most of the time the vector notation will be used, but we will not specify this as it should be clear from the context which notation is meant in each expression. Let $\|x\|$ and $\langle x, y \rangle$ denote the usual norm of $x \in \mathbb{R}^{\boldsymbol{n}}$ and the usual scalar product between $x, y \in \mathbb{R}^{\boldsymbol{n}}$, respectively.

Given any set S, $|S|$ denotes its cardinality; in particular, $|[\boldsymbol{k}]| = k_1 k_2$. We denote the matrices and operators by upright capital letters. The dimensions of matrices and multidimensional distributions should be clear from the context. For two positive sequences $(a_l), (b_l)$, $a_l \asymp b_l$ means $c^{-1} b_l \le a_l \le c b_l$ with some absolute $c > 0$.

For $\boldsymbol{k} \in [\boldsymbol{n}]$, consider a mapping $\boldsymbol{z} = (z_1, z_2) : [\boldsymbol{n}] \mapsto [\boldsymbol{k}]$, where $z_1 : [n_1] \mapsto [k_1]$ and $z_2 : [n_2] \mapsto [k_2]$. Denote $[\boldsymbol{k}]^{[\boldsymbol{n}]} = ([k_1]^{[n_1]}, [k_2]^{[n_2]})$, where $[k]^{[n]}$ is the collection of all surjective functions from $[n]$ to $[k]$ for $k \in [n]$. Each mapping $\boldsymbol{z} \in [\boldsymbol{k}]^{[\boldsymbol{n}]}$ uniquely determines the pertinent partition $\boldsymbol{I} = \boldsymbol{I}(\boldsymbol{z})$ of the rows and columns of any matrix $(M_{ij}) \in \mathbb{R}^{\boldsymbol{n}}$ into $|[\boldsymbol{k}]| = k_1 k_2$ blocks:

$$[\boldsymbol{n}] = ([n_1], [n_2]) = \boldsymbol{z}^{-1}([\boldsymbol{k}]) = \big(z_1^{-1}([k_1]), z_2^{-1}([k_2])\big) = \cup_{(I,J) \in \boldsymbol{I}} (I, J),$$

so that $\boldsymbol{I} = \boldsymbol{I}(\boldsymbol{z}) = \cup_{(i,j) \in [\boldsymbol{k}]} (I_i, J_j)$ is the corresponding partition, where $(I_i, J_j) = \big(I_i(z_1), J_j(z_2)\big) = \big(z_1^{-1}(i), z_2^{-1}(j)\big) \subseteq [\boldsymbol{n}]$. Thus, the collection of all mappings $\mathcal{Z} = \mathcal{Z}(\boldsymbol{n}) = \{\boldsymbol{z} \in [\boldsymbol{k}]^{[\boldsymbol{n}]}, \boldsymbol{k} \in [\boldsymbol{n}]\}$ yields the collection of all *biclustered* partitions of $[\boldsymbol{n}]$: $\mathcal{I}' = \mathcal{I}'(\boldsymbol{n}) = \big\{\boldsymbol{I}(\boldsymbol{z}), \boldsymbol{z} \in [\boldsymbol{k}]^{[\boldsymbol{n}]}, \boldsymbol{k} \in [\boldsymbol{n}]\big\}$. For each partition $\boldsymbol{I} \in \mathcal{I}'$ denote by $\boldsymbol{k}(\boldsymbol{I}) = (k_1(\boldsymbol{I}), k_2(\boldsymbol{I})) \in [\boldsymbol{n}]$ the numbers of nonempty row and column blocks in $\boldsymbol{I}$.

For each $\boldsymbol{I} \in \mathcal{I}'$, define the linear subspace of $\mathbb{R}^{\boldsymbol{n}}$

$$\mathbb{L}_{\boldsymbol{I}} = \big\{x \in \mathbb{R}^{\boldsymbol{n}} : x_{ij} = x_{i'j'} \ \forall\, (i,j), (i',j') \in (I,J),\ \forall\, (I,J) \in \boldsymbol{I}\big\}. \tag{4.2}$$

Clearly, if $\boldsymbol{k}(\boldsymbol{I}) = \boldsymbol{k}$, then $\dim(\mathbb{L}_{\boldsymbol{I}}) = |[\boldsymbol{k}]| = k_1 k_2$.

4.1.2. SLICING INTO COMPLEXITY LAYERS

We call $\boldsymbol{k}(\boldsymbol{I}) \in [\boldsymbol{n}]$ the *biclustering complexity* of partition $\boldsymbol{I}$ (sometimes we will use term *structure* $\boldsymbol{I}$). In a way, the quantity $\boldsymbol{k}(\boldsymbol{I})$ describes how complex the biclustering structure of the linear space $\mathbb{L}_{\boldsymbol{I}}$ is, so it will also be the biclustering complexity (or just complexity) of $\mathbb{L}_{\boldsymbol{I}}$.

The family of linear spaces $\{\mathbb{L}_{\boldsymbol{I}}, \boldsymbol{I} \in \mathcal{I}'\}$, defined by (4.2), leads to the slicing of the parameter space $\mathbb{R}^{\boldsymbol{n}} = \cup_{\boldsymbol{I}\in\mathcal{I}'}\mathbb{L}_{\boldsymbol{I}}$. The idea is to slice the entire space $\mathbb{R}^{\boldsymbol{n}}$ into "layers" of linear spaces with equal biclustering complexities. Let us unite all $\mathbb{L}_{\boldsymbol{I}}$ with the same biclustering complexity $\boldsymbol{k}$ in the *complexity layers* $\mathcal{L}_{\boldsymbol{k}} = \cup_{\boldsymbol{I}\in\mathcal{I}'_{\boldsymbol{k}}}\mathbb{L}_{\boldsymbol{I}}$, where the index set $\mathcal{I}'_{\boldsymbol{k}} = \mathcal{I}'_{\boldsymbol{k}}(\boldsymbol{n}) = \{\boldsymbol{I} \in \mathcal{I}' : \boldsymbol{k}(\boldsymbol{I}) = \boldsymbol{k}\}$ describes the family of all partitions (of $[\boldsymbol{n}]$) with biclustering complexity $[\boldsymbol{k}]$. This yields the slicing of the entire parameter space into complexity layers $\mathbb{R}^{\boldsymbol{n}} = \cup_{\boldsymbol{k}\in[\boldsymbol{n}]}\mathcal{L}_{\boldsymbol{k}}$.

All the linear spaces $\mathbb{L}_{\boldsymbol{I}}$ from the complexity layer $\mathcal{L}_{\boldsymbol{k}}$ have the common dimension $d_{\boldsymbol{k}} \triangleq \dim(\mathbb{L}_{\boldsymbol{I}}) = |[\boldsymbol{k}(\boldsymbol{I})]| = k_1 k_2$. This is one ingredient to characterize the (biclustering) complexity of the whole layer $\mathcal{L}_{\boldsymbol{k}}$. Yet another important ingredient of complexity of the layer $\mathcal{L}_{\boldsymbol{k}}$ is its cardinality $|\mathcal{L}_{\boldsymbol{k}}|$ which is the number of *different* spaces $\mathbb{L}_{\boldsymbol{I}}$ in $\mathcal{L}_{\boldsymbol{k}}$. Let us elucidate this.

Suppose the following condition is fulfilled for the noise ξ in (4.1): $\mathbb{E}_\theta \xi_{ij} = 0$, $(i,j) \in [\boldsymbol{n}]$; and for some $\alpha > 0$,

$$\mathbb{E}_\theta \exp\left\{\alpha \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\right\} \le \exp\{d_{\boldsymbol{k}}\} \quad \text{for all } \mathbb{L}_{\boldsymbol{I}} \in \mathcal{L}_{\boldsymbol{k}},\ \boldsymbol{k} \in [\boldsymbol{n}],\ \theta \in \mathbb{R}^{\boldsymbol{n}}. \tag{C0}$$

Using (C0) and Jensen's inequality, we derive for each $\boldsymbol{k} \in [\boldsymbol{n}]$

$$\begin{aligned}\exp\left\{\alpha \mathbb{E}_\theta \max_{\boldsymbol{I}\in\mathcal{I}'_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\right\} = \exp\left\{\alpha \mathbb{E}_\theta \max_{\mathbb{L}_{\boldsymbol{I}}\in\mathcal{L}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\right\} &\le \mathbb{E}_\theta \exp\left\{\alpha \max_{\mathbb{L}_{\boldsymbol{I}}\in\mathcal{L}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\right\} \\ &\le \sum_{\mathbb{L}_{\boldsymbol{I}}\in\mathcal{L}_{\boldsymbol{k}}} \mathbb{E}_\theta e^{\alpha\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2} \le e^{d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}|}. \end{aligned} \tag{4.3}$$

Hence, under (C0) we can control the maximal projected error $\max_{\mathbb{L}_{\boldsymbol{I}}\in\mathcal{L}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2$ up to the order of the quantity $d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}|$. It is this sum of the common dimension $d_{\boldsymbol{k}}$ of all $\mathbb{L}_{\boldsymbol{I}} \in \mathcal{L}_{\boldsymbol{k}}$ and the log of the cardinality of $\mathcal{L}_{\boldsymbol{k}}$ that characterizes the *total complexity* of the whole layer $\mathcal{L}_{\boldsymbol{k}}$. This is the reason why we use (a multiple of) $d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}|$ as penalty term in the penalization method and this quantity would also enter the local (oracle) rate. In fact, any majorant $\rho(\boldsymbol{k}) \ge C_1 d_{\boldsymbol{k}} + C_2 \log|\mathcal{L}_{\boldsymbol{k}}|$ (for some $C_1, C_2 > 0$) can be taken as penalty; see (4.9) and (4.12) below. This may be useful when the quantity $\log|\mathcal{L}_{\boldsymbol{k}}|$ is difficult to compute, whereas some closed form upper bound can easily be derived. Of course, this comes at the price of a bigger resulting local rate because then the majorant of the total complexity will enter the local rate.

Let us propose a reasonable majorant for the total layer complexity $d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}| = k_1 k_2 + \log|\mathcal{L}_{\boldsymbol{k}}|$. Since the number of partitions $\boldsymbol{I}$ in $\mathcal{I}'_{\boldsymbol{k}}$ is at least as big as the number of different linear spaces $\mathbb{L}_{\boldsymbol{I}}$ in $\mathcal{L}_{\boldsymbol{k}}$, we have that $|\mathcal{L}_{\boldsymbol{k}}| \le |\mathcal{I}'_{\boldsymbol{k}}|$. The cardinality of the set $\mathcal{I}'_{\boldsymbol{k}}$ of partitions of complexity $[\boldsymbol{k}] \in [\boldsymbol{n}]$ is $|\mathcal{I}'_{\boldsymbol{k}}| = N(n_1, k_1)N(n_2, k_2)$, where $N(n,k)$ is the number of ways to put n different objects into k different boxes so that each box contains at least one object. Notice that $S(n,k) = N(n,k)/k! = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$ is a Stirling number of the second kind. To have a simple closed form expression for a majorant of

the complexity, instead of $|\mathcal{I}'_{\boldsymbol{k}}|$ we can use its upper bound $|\bar{\mathcal{I}}_{\boldsymbol{k}}| = k_1^{n_1} k_2^{n_2}$, where $\bar{\mathcal{I}}_{\boldsymbol{k}} = \cup_{\boldsymbol{k}' \le \boldsymbol{k}} \mathcal{I}'_{\boldsymbol{k}'}$ ($\boldsymbol{k}' \le \boldsymbol{k}$ means that $k'_1 \le k_1$ and $k'_2 \le k_2$). Indeed, since $\mathcal{I}'_{\boldsymbol{k}} \subseteq \bar{\mathcal{I}}_{\boldsymbol{k}}$, $|\mathcal{I}'_{\boldsymbol{k}}| \le |\bar{\mathcal{I}}_{\boldsymbol{k}}| = k_1^{n_1} k_2^{n_2}$. We obtain the bound for the layer complexity: for $\boldsymbol{k} \in [\boldsymbol{n}]$,

$$d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}| \le d_{\boldsymbol{k}} + \log|\mathcal{I}'_{\boldsymbol{k}}| \le d_{\boldsymbol{k}} + \log|\bar{\mathcal{I}}_{\boldsymbol{k}}| = k_1 k_2 + n_1 \log k_1 + n_2 \log k_2. \tag{4.4}$$

Remark 4.1. *Notice that $d_{\boldsymbol{k}} + \log|\bar{\mathcal{I}}_{\boldsymbol{k}}|$ can be seen as an upper bound for the total complexity of the even more saturated layer $\bar{\mathcal{L}}_{\boldsymbol{k}} = \cup_{\boldsymbol{k}' \le \boldsymbol{k}} \mathcal{L}_{\boldsymbol{k}'}$. The corresponding family of partitions $\bar{\mathcal{I}}_{\boldsymbol{k}} = \cup_{\boldsymbol{k}' \le \boldsymbol{k}} \mathcal{I}'_{\boldsymbol{k}'}$ contains all the partitions of $[\boldsymbol{n}]$ into $\boldsymbol{k}$ blocks (some of which are possibly empty). Indeed, for each $\mathbb{L}_{\boldsymbol{I}} \in \bar{\mathcal{L}}_{\boldsymbol{k}}$, $\dim(\mathbb{L}_{\boldsymbol{I}}) + \log|\bar{\mathcal{L}}_{\boldsymbol{k}}| \le \dim(\mathbb{L}_{\boldsymbol{I}}) + \log|\bar{\mathcal{I}}_{\boldsymbol{k}}| \le \max_{\mathbb{L}_{\boldsymbol{I}} \in \bar{\mathcal{L}}_{\boldsymbol{k}}} \dim(\mathbb{L}_{\boldsymbol{I}}) + \log|\bar{\mathcal{I}}_{\boldsymbol{k}}| = d_{\boldsymbol{k}} + \log|\bar{\mathcal{I}}_{\boldsymbol{k}}| = k_1 k_2 + n_1 \log k_1 + n_2 \log k_2$ since $\max_{\mathbb{L}_{\boldsymbol{I}} \in \bar{\mathcal{L}}_{\boldsymbol{k}}} \dim(\mathbb{L}_{\boldsymbol{I}}) = k_1 k_2$ and $|\bar{\mathcal{I}}_{\boldsymbol{k}}| = k_1^{n_1} k_2^{n_2}$.*

4

4.1.3. "CLEANING UP" THE COMPLEXITY LAYERS

The problem with the bound (4.4) is that for some biclustering complexities $\boldsymbol{k} \in [\boldsymbol{n}]$ there may be too many different partitions $\boldsymbol{I}_1, \dots, \boldsymbol{I}_L \in \mathcal{I}'_{\boldsymbol{k}}$ corresponding to the same linear space, i.e., $\mathbb{L}_{\boldsymbol{I}_i} = \mathbb{L}_{\boldsymbol{I}_1}$, $i = 2, \dots, L$. Then the bound $|\mathcal{L}_{\boldsymbol{k}}| \le |\mathcal{I}'_{\boldsymbol{k}}|$ becomes too crude for those complexities $\boldsymbol{k}$. In particular, this bound is too crude for the cases (i) $\boldsymbol{k} \in \mathcal{K}_1 = \{\boldsymbol{k} \in [\boldsymbol{n}] : k_1 < n_1, k_2 = n_2\}$, (ii) $\boldsymbol{k} \in \mathcal{K}_2 = \{\boldsymbol{k} \in [\boldsymbol{n}] : k_1 = n_1, k_2 < n_2\}$, and (iii) $\boldsymbol{k} \in \mathcal{K}_3 = \{(n_1, n_2)\}$. Indeed, let $\mathrm{id}_m : [m] \mapsto [m]$ with $\mathrm{id}_m(k) = k$, $k \in [m]$, the identity mapping of $[m]$. Then it is easy to see that $\mathbb{L}_{\boldsymbol{I}(z_1,z_2)} = \mathbb{L}_{\boldsymbol{I}(z_1,\mathrm{id}_{n_2})}$ for all $z_2 \in [n_2]^{[n_2]}$ and all $z_1 \in [k_1]^{[n_1]}$, $k_1 \in [n_1]$. Similarly, $\mathbb{L}_{\boldsymbol{I}(z_1,z_2)} = \mathbb{L}_{\boldsymbol{I}(\mathrm{id}_{n_1},z_2)}$ for all $z_1 \in [n_1]^{[n_1]}$, $z_2 \in [k_2]^{[n_2]}$, $k_2 \in [n_2]$; and $\mathbb{L}_{\boldsymbol{I}(z_1,z_2)} = \mathbb{L}_{\boldsymbol{I}(\mathrm{id}_{n_1},\mathrm{id}_{n_2})}$ for all $\boldsymbol{z} \in [\boldsymbol{n}]^{\boldsymbol{n}}$. Hence, $|\mathcal{L}_{\boldsymbol{k}}| \le \left|[k_1]^{[n_1]}\right| \le k_1^{n_1}$ for $\boldsymbol{k} \in \mathcal{K}_1$, $|\mathcal{L}_{\boldsymbol{k}}| \le k_2^{n_2}$ for $\boldsymbol{k} \in \mathcal{K}_2$, and $|\mathcal{L}_{\boldsymbol{k}}| \le 1$ for $\boldsymbol{k} \in \mathcal{K}_3$. Thus, we improve the bound (4.4) by giving the following majorant for the total complexity of the layer $\mathcal{L}_{\boldsymbol{k}}$: $d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}| \le \rho(\boldsymbol{k})$, $\boldsymbol{k} \in [\boldsymbol{n}]$, where $\rho(\boldsymbol{k}) = \rho_{\boldsymbol{n}}(\boldsymbol{k})$ is defined as follows:

$$\rho(\boldsymbol{k}) \triangleq \begin{cases} k_1 k_2 + n_1 \log k_1 + n_2 \log k_2, & k_1 < n_1,\ k_2 < n_2, \\ k_1 n_2 + n_1 \log k_1, & k_1 < n_1,\ k_2 = n_2, \\ n_1 k_2 + n_2 \log k_2, & k_1 = n_1,\ k_2 < n_2, \\ n_1 n_2, & k_1 = n_1,\ k_2 = n_2. \end{cases} \tag{4.5}$$

Recall the collection $\mathcal{I}'$ of all the partitions $\mathcal{I}' = \cup_{\boldsymbol{k} \in [\boldsymbol{n}]} \mathcal{I}'_{\boldsymbol{k}}$, where $\mathcal{I}'_{\boldsymbol{k}} = \mathcal{I}'_{\boldsymbol{k}}(\boldsymbol{n}) = \{\boldsymbol{I} \in \mathcal{I}' : \boldsymbol{k}(\boldsymbol{I}) = \boldsymbol{k}\}$ is the collection of all possible partitions of $[\boldsymbol{n}]$ into nonempty $\boldsymbol{k}$ blocks. It is clear from the above discussion that there is no need to consider all the partitions $\mathcal{I}'$, but only a subset $\mathcal{I} \subset \mathcal{I}'$ of those $\boldsymbol{I} \in \mathcal{I}$ which generate all different linear spaces $\{\mathbb{L}_{\boldsymbol{I}}, \boldsymbol{I} \in \mathcal{I}'\}$. Now we "clean up" the original collection of all the partitions $\mathcal{I}' = \cup_{\boldsymbol{k} \in [\boldsymbol{n}]} \mathcal{I}'_{\boldsymbol{k}}$ by removing redundant partitions from it. Consider a family $\mathcal{I}'_{\boldsymbol{k}}$ and denote $l_{\boldsymbol{k}} = |\mathcal{L}_{\boldsymbol{k}}|$. Then $\mathcal{L}_{\boldsymbol{k}} = \cup_{\boldsymbol{I} \in \mathcal{I}'_{\boldsymbol{k}}} \mathbb{L}_{\boldsymbol{I}} = \cup_{i=1}^{l_{\boldsymbol{k}}} \mathbb{L}_{\boldsymbol{I}_i}$ for some different linear spaces $\mathbb{L}_{\boldsymbol{I}_i}$, $i = 1, \dots, l_{\boldsymbol{k}}$ (different choice for $\boldsymbol{I}_i$'s are possible, take any one). Now define the cleaned up version of $\mathcal{I}'_{\boldsymbol{k}}$ as $\mathcal{I}_{\boldsymbol{k}} = \{\mathbb{L}_{\boldsymbol{I}_i}, i = 1, \dots, l_{\boldsymbol{k}}\}$. Clearly, $|\mathcal{L}_{\boldsymbol{k}}| = |\mathcal{I}_{\boldsymbol{k}}|$. By doing this for each $\mathcal{I}'_{\boldsymbol{k}}$, $\boldsymbol{k} \in [\boldsymbol{n}]$, we obtain the new family of partitions (which we will use from now on)

$$\mathcal{I} = \cup_{\boldsymbol{k} \in [\boldsymbol{n}]} \mathcal{I}_{\boldsymbol{k}}. \tag{4.6}$$

Since $\mathcal{I}_{\boldsymbol{k}} \subseteq \mathcal{I}'_{\boldsymbol{k}}$, $\mathcal{I} \subseteq \mathcal{I}'$, i.e., the family $\mathcal{I}$ is a thinned version of the original family $\mathcal{I}'$ of all partitions. But this family still covers all possible biclustering structures for the parameter $\theta \in \mathbb{R}^{\boldsymbol{n}}$ as we retain the slicing of the entire parameter space into complexity layers in the sense that $\mathbb{R}^{\boldsymbol{n}} = \cup_{\boldsymbol{k} \in [\boldsymbol{n}]} \mathcal{L}_{\boldsymbol{k}}$, with $\mathcal{L}_{\boldsymbol{k}} = \cup_{\boldsymbol{I} \in \mathcal{I}'_{\boldsymbol{k}}} \mathbb{L}_{\boldsymbol{I}} = \cup_{\boldsymbol{I} \in \mathcal{I}_{\boldsymbol{k}}} \mathbb{L}_{\boldsymbol{I}}$, $\boldsymbol{k} \in [\boldsymbol{n}]$.

Remark 4.2. *An even better majorant for the total complexity of the layer $\mathcal{L}_{\boldsymbol{k}}$ is the complexity itself $\rho_0(\boldsymbol{k}) = d_{\boldsymbol{k}} + \log|\mathcal{L}_{\boldsymbol{k}}| = d_{\boldsymbol{k}} + \log|\mathcal{I}_{\boldsymbol{k}}|$. But the analytic expression for $|\mathcal{I}_{\boldsymbol{k}}| = |\mathcal{L}_{\boldsymbol{k}}|$ is not available. So we will use the upper bound* (4.5).

Remark 4.3. *One can see the resulting family $\mathcal{I}$ of partitions as equivalence classes on the original collection of all partition $\mathcal{I}'$ with the equivalence relation: $\boldsymbol{I}_1 \sim \boldsymbol{I}_2$ if and only if $\mathbb{L}_{\boldsymbol{I}_1} = \mathbb{L}_{\boldsymbol{I}_2}$.*

Remark 4.4. *There is redundancy in the slicing $\mathbb{R}^{\boldsymbol{n}} = \cup_{\boldsymbol{k} \in [\boldsymbol{n}]} \mathcal{L}_{\boldsymbol{k}}$ as some of the spaces are subspaces of others. This redundancy is not an issue to some extent, namely, as long as the cardinality of redundant pieces is not too big (to affect the right order of complexity).*

4

4.1.4. CONDITION ON THE ERROR

In the foregoing, we proposed condition (C0) which in turn leads to the definition (4.5) of the complexity of layer $\mathcal{I}_{\boldsymbol{k}}$. We can actually relax condition (C0) by allowing $d_{\boldsymbol{k}} + \log|\mathcal{I}_{\boldsymbol{k}}|$ instead of just $d_{\boldsymbol{k}}$ in the exponent of the right hand side of the condition. Precisely, in this chapter we impose the following condition, which will be assumed throughout.

CONDITION (C1). The random variables $\xi = (\xi_{ij})$ in (4.1) satisfy: $\mathbb{E}_\theta \xi_{ij} = 0$, $(i,j) \in [\boldsymbol{n}]$; and for some $\alpha > 0$ (without loss of generality assume $\alpha \in (0,1]$),

$$\mathbb{E}_\theta \exp\left\{\alpha \|\mathrm{P}_{\boldsymbol{I}} \xi\|^2\right\} \le e^{\rho(\boldsymbol{k}(\boldsymbol{I}))} \quad \text{for all } \boldsymbol{I} \in \mathcal{I}, \tag{C1}$$

where $\rho(\boldsymbol{k})$ is defined by (4.5).

Notice that the unknown distribution of ξ may also depend on θ, in that case we assume Condition (C1) to be fulfilled for all $\theta \in \mathbb{R}^{\boldsymbol{n}}$. The constant $\alpha \in (0,1]$ will be fixed in the sequel and we omit the dependence on this constant in all further notation. Typically (C1) holds with just the first term $d_{\boldsymbol{k}(\boldsymbol{I})}$ of $\rho(\boldsymbol{k}(\boldsymbol{I}))$ instead of the whole $\rho(\boldsymbol{k}(\boldsymbol{I}))$ (see examples below), but we can let Condition (C1) be slightly weaker without changing the proof, at the expense of slightly bigger constants.

Note that in view of Remark 1.6 Condition (C1) is satisfied for independent normal ξ_{ij}'s. Let us show that Condition (C1) is satisfied for bounded ξ_{ij}'s. We should also mention that in case of independent normal or Bernoulli errors some constants in the proofs can be sharpened.

In case $X_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$, we have $\xi_{ij} \in \{1-\theta_{ij}, -\theta_{ij}\} \subseteq [-1,1]$ and $\mathbb{E}_\theta \xi_{ij} = 0$, $(i,j) \in [\boldsymbol{n}]$. We can represent the projection $\mathrm{P}_{\boldsymbol{I}} = \mathrm{B}\mathrm{B}^T$, where B is the $(n_1 n_2 \times k_1 k_2)$-matrix whose columns $(b_{IJ}, (I,J) \in \boldsymbol{I})$ form an orthonormal basis of $\mathbb{L}_{\boldsymbol{I}}$. Then $\|\mathrm{P}_{\boldsymbol{I}} \xi\|^2 = \|\mathrm{B}^T \xi\|^2 = \|\eta\|^2$, where $\eta = (\eta_{IJ}, (I,J) \in \boldsymbol{I})$, $\eta_{IJ} = b_{IJ}^T \xi$. We choose the following orthogonal basis of $\mathbb{L}_{\boldsymbol{I}}$: $b_{IJ} = \big((|I||J|)^{-1/2} 1\{(i,j) \in (I,J)\}, (i,j) \in [\boldsymbol{n}]\big)$, $(I,J) \in \boldsymbol{I}$, so that $\eta_{IJ} = \frac{1}{\sqrt{|I||J|}} \sum_{(i,j) \in (I,J)} \xi_{ij}$. Hoeffding's inequality implies that for any $t \ge 0$

$$\mathbb{P}_\theta(|\eta_{IJ}| \ge t) \le 2e^{-t^2/2}, \quad \text{for all } (I,J) \in \boldsymbol{I}.$$

Using this, we obtain for any $0 < b < 1/2$

$$\mathbb{E}_\theta e^{b\eta_{IJ}^2} = 1 + \int_1^\infty \mathbb{P}_\theta\big(e^{b\eta_{IJ}^2} \ge t\big)\,dt \le 1 + 2\int_1^\infty e^{-(\log t)/(2b)}\,dt = 1 + \frac{4b}{1-2b}.$$

By taking $b = 1/6$, we derive $\mathbb{E}_\theta \exp\{\frac{1}{6}\|\mathrm{P}_I\xi\|^2\} = \mathbb{E}_\theta \exp\{\frac{1}{6}\|\eta\|^2\} \le 2^{d_{\boldsymbol{k}(I)}} \le e^{d_{\boldsymbol{k}(I)}}$, which is Condition (C1) with the constant $\alpha = 1/6$ (in fact, the stronger condition (C0) is fulfilled). Of course, the above argument applies (with minor adjustments) to any independent zero mean bounded errors $\xi_{ij} \in [-c, c]$ for some $c > 0$.

In the proof of Theorem 4.1 below, we will need a bound for $\big[\mathbb{E}_\theta\|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big]^{1/2}$, for each $\boldsymbol{I} \in \mathcal{I}$. Actually, Condition (C1) ensures such a bound. Indeed, since $x^2 \le e^{2x}$ for all $x \ge 0$, by using the Hölder inequality and (C1), we derive that for any $t \in (0, \alpha]$ and $\boldsymbol{I} \in \mathcal{I}$,

$$\mathbb{E}_\theta\|\mathrm{P}_{\boldsymbol{I}}\xi\|^4 = \frac{4}{t^2}\mathbb{E}_\theta\big(\tfrac{t}{2}\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\big)^2 \le \frac{4}{t^2}\mathbb{E}_\theta e^{t\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2} \le \frac{4}{t^2}\big[\mathbb{E}_\theta e^{\alpha\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2}\big]^{t/\alpha} \le \frac{4}{t^2}e^{t\alpha^{-1}\rho(\boldsymbol{k}(\boldsymbol{I}))}.$$

To summarize, Condition (C1) implies that for any $t \in (0, 1/2]$

$$\big[\mathbb{E}_\theta\|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big]^{1/2} \le \frac{1}{\alpha t}\exp\big\{t\rho(\boldsymbol{k}(\boldsymbol{I}))\big\}, \quad \text{for all } \boldsymbol{I} \in \mathcal{I}. \tag{4.7}$$

In case $\xi_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, since $\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 \sim \chi^2_{d_{\boldsymbol{k}(\boldsymbol{I})}}$, in the proof of Theorem 4.1 we can use a better bound $\big[\mathbb{E}\|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big]^{1/2} = \big(d^2_{\boldsymbol{k}(\boldsymbol{I})} + 2d_{\boldsymbol{k}(\boldsymbol{I})}\big)^{1/2} \le d_{\boldsymbol{k}(\boldsymbol{I})} + 1 \le \rho(\boldsymbol{k}(\boldsymbol{I})) + 1$ instead of (4.7).

It is interesting to relate Condition (C1) to the so called *sub-gaussianity* condition on the error vector ξ. The random vector ξ is called *sub-gaussian* with parameter $\rho > 0$ if

$$\mathbb{P}(|\langle v, \xi\rangle| > t) \le e^{-\rho t^2} \quad \text{for all } t \ge 0 \text{ and all } v \in \mathbb{R}^{\boldsymbol{n}} \text{ such that } \|v\| = 1. \tag{4.8}$$

The sub-gaussianity condition (4.8) and Condition (C1) are close, but in general incomparable. For example, let $\xi_i = \xi_0$, $i \in [n]$, for some bounded random variable ξ_0 (say, uniform on $[-1, 1]$), then Condition (C1) trivially holds whereas the sub-gaussianity condition is not fulfilled. It is easy to see that the sub-gaussianity condition is equivalent to Condition (C1) for independent ξ_i's.

However, note that the ξ_i's are not assumed to follow any specific distribution and do not have to be even independent. In a way, Condition (C1) prevents too much dependence, but it still allows some interesting cases of dependent ξ_{ij}'s. For example, in the last version of the Arxiv-preprint [13] we showed that Condition (C1) is fulfilled for the vector ξ whose coordinates follow an autoregressive model.

4.1.5. PENALIZATION METHOD

Introduce the penalized structure selector $\hat{\boldsymbol{I}}$ (and the corresponding selector $\hat{\boldsymbol{k}} = \boldsymbol{k}(\hat{\boldsymbol{I}})$ of numbers of row and column blocks) by minimizing (if not unique, take any minimizer) the penalized criterion $\|X - \mathrm{P}_{\boldsymbol{I}}X\|^2 + K\rho(\boldsymbol{k}(\boldsymbol{I}))$, i.e.,

$$\hat{\boldsymbol{I}} = \operatorname*{argmin}_{\boldsymbol{I}\in\mathcal{I}}\big\{\|X - \mathrm{P}_{\boldsymbol{I}}X\|^2 + K\rho(\boldsymbol{k}(\boldsymbol{I}))\big\}, \quad \hat{\boldsymbol{k}} = \boldsymbol{k}(\hat{\boldsymbol{I}}), \tag{4.9}$$

where $K > 0$ is a penalization constant, the family $\mathcal{I}$ is given by (4.6) and the complexity $\rho(\boldsymbol{k})$ is defined by (4.5). Instead of $\mathcal{I}$, one can use the bigger family $\mathcal{I}'$ of all biclustering partitions, since the minimum of the criterion does not change.

The method (4.9) is nothing else but the penalization method with the complexity penalty $K\rho(\boldsymbol{k}(\boldsymbol{I}))$. Plugging in the penalized structure selector $\hat{\boldsymbol{I}}$ into the projection operator $\mathrm{P}_{\boldsymbol{I}}X$ yields the corresponding penalized projection estimator for θ:

$$\hat{\theta} = \hat{\theta}(\hat{\boldsymbol{I}}) = \mathrm{P}_{\hat{\boldsymbol{I}}}X. \tag{4.10}$$

Later, in the proofs we will need the series $\sum_{\boldsymbol{I}\in\mathcal{I}} \exp\{-K\rho(\boldsymbol{k}(\boldsymbol{I}))\}$ to be bounded uniformly in $\boldsymbol{n}$. This is provided by assuming $K \geq 1$. Indeed, since $\rho(\boldsymbol{k}) \geq k_1 k_2 + \log|\mathcal{I}_{\boldsymbol{k}}|$, we derive for any $K \geq 1$ that

$$\begin{aligned}\sum_{\boldsymbol{I}\in\mathcal{I}} \exp\{-K\rho(\boldsymbol{k}(\boldsymbol{I}))\} &= \sum_{\boldsymbol{k}\in[\boldsymbol{n}],\boldsymbol{I}\in\mathcal{I}_{\boldsymbol{k}}} \exp\{-K\rho(\boldsymbol{k})\} = \sum_{\boldsymbol{k}\in[\boldsymbol{n}]} |\mathcal{I}_{\boldsymbol{k}}| \exp\{-K\rho(\boldsymbol{k})\} \\ &\leq \sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_2} k_1^{n_1} k_2^{n_2} e^{-K\rho(\boldsymbol{k})} \leq \sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_2} e^{-Kk_1k_2} \\ &= (e^K + e^{-K} - 2)^{-1} \triangleq \Sigma(K). \end{aligned} \tag{4.11}$$

Besides, in the proofs we will need the following technical condition on the penalization constant ensuring basically that the penalization constant K is large enough.

CONDITION (C2). With constant α from Condition (C1),

$$K \geq \bar{K} \triangleq \tfrac{1}{2} + \tfrac{21}{\alpha}. \tag{C2}$$

Remark 4.5. *Notice that Condition (C2) implies of course that $K \geq 1$ as $\alpha \in (0,1]$ so that (4.11) holds under (C2). It is possible to relax Condition (C2) by having a smaller number $\bar{K}$, if we use more precise (but significantly longer) evaluations in the proofs, but some nontrivial lower bound $\bar{K} \geq 1$ is unavoidable. Actually, a lower bound on the penalization constant is typical for penalization methods, an insightful discussion on this issue can be found in [23] (for the normal means problem). Basically, a lower bound on the penalization constant ensures that the complexity term is sufficiently prominent in the criterion, which is in line with the general principle of balancing model fit against model complexity.*

4.1.6. ORACLE CONVERGENCE RATE

For $\boldsymbol{I} \in \mathcal{I}$, consider the projection estimator $\mathrm{P}_{\boldsymbol{I}}X$ for estimating θ. Exactly in the same way as (4.3) follows from (C0), condition (C1) entails that $\mathbb{E}_\theta \max_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}(\boldsymbol{I})}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2 \leq C\rho(\boldsymbol{k}(\boldsymbol{I}))$ for some $C > 0$, where the complexity $\rho(\boldsymbol{k})$ is defined by (4.5). We immediately obtain the following upper bound for the estimator $\mathrm{P}_{\boldsymbol{I}}X$: for some $C > 0$,

$$\begin{aligned}\mathbb{E}_\theta\|\theta - \mathrm{P}_{\boldsymbol{I}}X\|^2 &\leq \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \mathbb{E}_\theta \max_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}(\boldsymbol{I})}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2 \\ &\leq \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + C\rho(\boldsymbol{k}(\boldsymbol{I})).\end{aligned}$$

This motivates the following definition. Introduce the family of local rates

$$r^2(\boldsymbol{I},\theta) = \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \rho(\boldsymbol{k}(\boldsymbol{I})), \quad \boldsymbol{I} \in \mathcal{I}.$$

For each $\theta \in \mathbb{R}^{\boldsymbol{n}}$, there exists the best choice of structure $\boldsymbol{I}_o = \boldsymbol{I}_o(\theta)$ (if not unique, take any minimizer) corresponding to the fastest rate from the family $\{r^2(\boldsymbol{I},\theta), \boldsymbol{I} \in \mathcal{I}\}$:

$$r^2(\theta) = \min_{\boldsymbol{I}\in\mathcal{I}} r^2(\boldsymbol{I},\theta) = r^2(\boldsymbol{I}_o,\theta) = \|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2 + \rho(\boldsymbol{k}(\boldsymbol{I}_o)), \tag{4.12}$$

representing the optimal trade-off between the approximation term $\|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2$ and the total complexity $\rho(\boldsymbol{k}(\boldsymbol{I}_o))$ of the *oracle structure* $\boldsymbol{I}_o$. From now on, we call the quantity $r^2(\theta)$ by *oracle rate*, and $\boldsymbol{k}_o = \boldsymbol{k}(\boldsymbol{I}_o)$ by *oracle complexity*. Note that $\boldsymbol{k}_o = \boldsymbol{k}_o(\theta)$ depends on θ since $\boldsymbol{I}_o$ depends on θ. Sometimes we will call the pair $(\boldsymbol{I}_o, \boldsymbol{k}_o) = (\boldsymbol{I}_o(\theta), \boldsymbol{k}_o(\theta))$ by *oracle* (or *oracle structure*).

Let us determine the range of values for the oracle rate $r^2(\theta)$. Recall the notation $N_{\boldsymbol{n}} \triangleq n_1 n_2$, the total number of observations. Let $\boldsymbol{I}_*$ be such that $\mathbb{L}_{\boldsymbol{I}_*} = \mathbb{R}^{\boldsymbol{n}}$ (the case of no structure), then $\boldsymbol{k}(\boldsymbol{I}_*) = \boldsymbol{n}$ and $\mathrm{P}_{\boldsymbol{I}_*}\theta = \theta$ for all $\theta \in \mathbb{R}^{\boldsymbol{n}}$. This and (4.5) imply the following trivial upper bound for the oracle rate (4.12):

$$r^2(\theta) = r^2(\boldsymbol{I}_o,\theta) \le r^2(\boldsymbol{I}_*,\theta) = \rho(\boldsymbol{k}(\boldsymbol{I}_*)) = \rho(\boldsymbol{n}) = n_1 n_2 = N_{\boldsymbol{n}}, \quad \theta \in \mathbb{R}^{\boldsymbol{n}}. \tag{4.13}$$

Next, recall the oracle complexity $\boldsymbol{k}_o(\theta) = \boldsymbol{k}(\boldsymbol{I}_o(\theta)) = (k_{o,1}(\theta), k_{o,2}(\theta))$, and define the set

$$\tilde{\Theta} = \{\theta \in \mathbb{R}^{\boldsymbol{n}} : \min(k_{o,1}(\theta), k_{o,2}(\theta)) = 1\}. \tag{4.14}$$

This is a "thin" subset of $\mathbb{R}^{\boldsymbol{n}}$ consisting of "highly structured parameters", whose oracle number of either row or block columns is 1. From (4.5), (4.12) and (4.13), it is easy to see that

$$N_{\boldsymbol{n}} \ge r^2(\theta) \ge \rho(\boldsymbol{k}_o(\theta)) \ge c N_{\boldsymbol{n}}^{1/2}, \quad \text{for } \theta \in \mathbb{R}^{\boldsymbol{n}} \backslash \tilde{\Theta}, \tag{4.15}$$

where we can take $c = \log 2$. We derived that the oracle rate $r^2(\theta)$ is sandwiched between $N_{\boldsymbol{n}}^{1/2}$ and $N_{\boldsymbol{n}}$ for the "majority" of θ's in the sense that $c N_{\boldsymbol{n}}^{1/2} \le r^2(\theta) \le N_{\boldsymbol{n}}$ for all $\theta \in \mathbb{R}^{\boldsymbol{n}} \backslash \tilde{\Theta}$.

4.1.7. ORACLE AND TRUE STRUCTURES

Besides the oracle structure $(\boldsymbol{I}_o, \boldsymbol{k}_o)$ defined above, one can consider the so called "true" structure of θ. Recall that if $\theta \in \mathbb{L}_{\boldsymbol{I}}$ for some $\boldsymbol{I} \in \mathcal{I}$ we say that θ has structure $\boldsymbol{I}$. Clearly, θ can have many structures, the true structure of θ is the one with the smallest complexity. Precisely, for each $\theta \in \mathbb{R}^{\boldsymbol{n}}$, define its underlying structure $\boldsymbol{I}^* = \boldsymbol{I}^*(\theta) \in \mathcal{I}$ as the one satisfying $\rho(\boldsymbol{k}(\boldsymbol{I}^*(\theta)) = \min\{\rho(\boldsymbol{k}(\boldsymbol{I})) : \theta \in \mathbb{L}_{\boldsymbol{I}}, \boldsymbol{I} \in \mathcal{I}\}$ (if not unique, take any minimizer). Then $\boldsymbol{k}^* = \boldsymbol{k}^*(\theta) = \boldsymbol{k}(\boldsymbol{I}^*(\theta)) = (k_1^*(\theta), k_2^*(\theta))$ has the meaning of "true" biclustering complexity of θ. The smaller the total complexity $\rho(\boldsymbol{k}^*(\theta))$, the more structured θ is. The "most structured case" is $\dim(\mathbb{L}_{\boldsymbol{I}^*}) = 1$ which happens if $\boldsymbol{k}^* = (1,1)$, i.e., the θ_{ij}'s are all in one block; the least structured case is $\dim(\mathbb{L}_{\boldsymbol{I}^*}) = \dim(\mathbb{R}^{\boldsymbol{n}}) = n_1 n_2$, which happens if $\boldsymbol{k}^* = \boldsymbol{n} = (n_1, n_2)$, i.e., there are $n_1 n_2$ blocks, each θ_{ij} occupying one block.

We say that the parameter θ has zero deceptiveness if the oracle recovers the true structure $(\boldsymbol{I}_o(\theta), \boldsymbol{k}_o(\theta)) = (\boldsymbol{I}^*(\theta), \boldsymbol{k}^*(\theta))$. In this case, the oracle rate only consists of complexity term: $r^2(\theta) = r^2(\boldsymbol{I}^*,\theta) = \|\theta - \mathrm{P}_{\boldsymbol{I}^*}\theta\|^2 + \rho(\boldsymbol{k}^*) = \rho(\boldsymbol{k}^*)$, because $\mathrm{P}_{\boldsymbol{I}^*}\theta = \theta$ $(\theta \in \mathbb{L}_{\boldsymbol{I}^*})$ by the definition of the true structure. The oracle rate is always not larger than the rate at the true structure: $r^2(\theta) = \|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2 + \rho(\boldsymbol{k}_o) \le r^2(\boldsymbol{I}^*,\theta) = \rho(\boldsymbol{k}^*)$ for all $\theta \in \mathbb{R}^{\boldsymbol{n}}$ by the

definition of the oracle. Therefore, it is always better to use the oracle structure than the true structure for the estimation problem.

Speaking informally, if $\|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2 > 0$, the corresponding θ is deceptive, the oracle is "tricked" by this deceptive θ and makes an approximation error $\|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2$. Interestingly, the oracle makes errors of only one type: the oracle can only "overstructure" (less blocks is more structure) and never "understructure", as the oracle can only merge some blocks of the true structure $\boldsymbol{I}^*$, leading to a smaller total complexity $\rho(\boldsymbol{k}_o) < \rho(\boldsymbol{k}^*)$ if $\|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2 > 0$.

4.2. MAIN RESULTS

In this section we give the main results of Chapter 4.

4.2.1. ORACLE ESTIMATION

The following theorem establishes that the penalized estimator $\hat{\theta}$, defined by (4.10), converges to θ with the oracle rate $r(\theta)$, uniformly over the entire parameter space $\mathbb{R}^{\boldsymbol{n}}$.

Theorem 4.1. *Let Conditions (C1) and (C2) be fulfilled. Then there exist positive constants M_1, H_1 and m_1 such that for any $\theta \in \mathbb{R}^{\boldsymbol{n}}$ and any $M \geq 1$,*

$$\mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 \geq M_1 r^2(\theta) + M\big) \leq H_1 e^{-m_1 M}.$$

A few remarks on the theorem are in order.

Remark 4.6. *The constants in the theorem depend only on α and some also on the penalization constant K, the exact expressions can be found in the proof.*

Remark 4.7. *In the beginning of this chapter we set $\sigma^2 = 1$ in the model (4.1) without loss of generality. Indeed, if $\sigma^2 \neq 1$, we first divide the model (4.1) by σ to get the new model $X'_{ij} = \theta'_{ij} + \xi_{ij}$ with $X'_{ij} = X_{ij}/\sigma$ and $\theta'_{ij} = \theta_{ij}/\sigma$. Since $\sigma^2 = 1$ in this new model, we have the results (Theorem 4.1 above and later other results) for θ'_{ij} in terms of X'_{ij}'s. Finally, we use these results for θ'_{ij}'s in terms of X'_{ij}'s to express the results for θ_{ij} in terms of X_{ij}'s.*

For example, in case $\sigma^2 \neq 1$, the penalization method (4.9) becomes $\hat{\boldsymbol{I}} = \operatorname{argmin}_{\boldsymbol{I}\in\mathcal{I}}\{\|X - \mathrm{P}_{\boldsymbol{I}}X\|^2 + K\sigma^2\rho(\boldsymbol{k}(\boldsymbol{I}))\}$, the oracle rate $r^2(\theta) = \min_{\boldsymbol{I}\in\mathcal{I}}\{\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \sigma^2\rho(\boldsymbol{k}(\boldsymbol{I}))\}$, and the claim of Theorem 4.1 reads as $\mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 \geq M_1 r^2(\theta) + \sigma^2 M\big) \leq H_1 e^{-m_1 M}$.

Remark 4.8. *The above non-asymptotic exponential bound in the theorem for the probability of deviation of $\|\hat{\theta} - \theta\|^2$ from a multiple of the oracle rate provides a rather refined characterization of the quality of the estimator $\hat{\theta}$, finer than typical oracle inequalities in expectation as, e.g., in [23] and [5]. Clearly, exponential probability bounds imply the traditional bounds in expectation by integration. This refined formulation also allows to easily study various asymptotic regimes: $n_1 \to \infty$, $n_2 \to \infty$; $\sigma \to 0$ (in case $\sigma^2 \neq 1$ in the model (4.1)); or their combination, as we can let M depend in any way on n_1, n_2, σ.*

For example, take $M = r^2(\theta)$ in the claim of the theorem, then we have $\mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 \geq (M_1 + 1)r^2(\theta)\big) \leq H_1 e^{-m_1 r^2(\theta)} \leq H_1 e^{-m_1\rho(\boldsymbol{k}_o)} \to 0$ if $\rho(\boldsymbol{k}_o) \to \infty$, which happens if $n_1 \to \infty$ and $k_{o,1} > 1$ or if $n_2 \to \infty$ and $k_{o,2} > 1$.

Remark 4.9. *As is discussed in Section 4.1.4, we could take a sharper quantity as majorant for* $\log|\mathcal{L}_{\boldsymbol{k}}|$ *in the definition (4.5) of* $\rho(\boldsymbol{k})$. *Namely, we can take* $\log N(n,k)$ *instead of* $n\log k$ *everywhere in the definition (4.5) of* $\rho(\boldsymbol{k})$, *obtaining a new majorant* $\rho_1(\boldsymbol{k})$. *Here* $N(n,k)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$ *is the number of ways to put n different objects into k different boxes so that each box contains at least one object (*$N(n,k)=k!S(n,k)$, *where* $S(n,k)$ *is the Stirling number of the second kind). Since* $N(n,k)\le k^n$, $\rho_1(\boldsymbol{k})\le\rho(\boldsymbol{k})$ *and using the majorant* $\rho_1(\boldsymbol{k})$ *instead of* $\rho(\boldsymbol{k})$ *as penalty in method (4.9) would lead to a smaller oracle rate (with* $\rho_1(\boldsymbol{k})$ *instead of* $\rho(\boldsymbol{k})$*) and hence stronger results in Theorem 4.1.*

When proving Theorem 4.1, we also obtain in passing a result about the the penalized structure selector $\hat{\boldsymbol{I}}$, saying basically that $\hat{\boldsymbol{I}}$ "lives" on a set that is, in a sense, almost as good as the oracle structure $\boldsymbol{I}_o$.

4

Theorem 4.2. *Let Conditions (C1) and (C2) be fulfilled,* c_1,c_2,c_3 *be the constants defined in Lemma 4.2,* $\Sigma(c_1)=(e^{c_1}+e^{-c_1}-2)^{-1}$ *be defined by (4.11). Then for any* $\theta\in\mathbb{R}^{\boldsymbol{n}}$ *and* $M\ge 0$,

$$\mathbb{P}_\theta\big(r^2(\hat{\boldsymbol{I}},\theta)\ge c_3 r^2(\theta)+M\big)\le\Sigma(c_1)e^{-c_2M}.$$

We can interpret the above result as some sort of *structure recovery*, or, if we put the problem in network context, *community detection* (see Section 4.3.2 about the stochastic block model), but in a somewhat weak sense. Namely, Theorem 4.2 says basically that the data based structure selector $\hat{\boldsymbol{I}}$ "lives" in the set of structures resembling the *oracle structure* $\boldsymbol{I}_o$ (notice that, in general the oracle structure is not the same as the so called true structure; see Section 4.1.7) in the sense that the rates for these structures are not too far above the oracle rate.

4.2.2. CONFIDENCE BALL

Theorem 4.1 establishes the strong local optimal properties of the estimator $\hat\theta$, but this does not solve the uncertainty quantification problem yet. For that, we need to construct a confidence ball and address its coverage and size properties. The estimator $\hat\theta$ is a good candidate for the center of the confidence ball. As to the choice of a data dependent radius, a first rough idea is to estimate the quantity $\|\theta-\hat\theta\|^2$ by $\|X-\hat\theta\|^2$ and use this as the quadratic radius of the confidence ball. However, it is clear that there is a lot of bias in $\|X-\hat\theta\|^2$, the biggest part of which is due to the term $\|\xi\|^2$. To de-bias for that part, we need to subtract its expectation. However, for typical error vectors ξ, even the de-biased quantity $\|\xi\|^2-\mathbb{E}\|\xi\|^2$ can only be controlled up to a margin of the order $|\boldsymbol{n}|^{1/2}=\sqrt{n_1n_2}$. That is why a term of the order $(n_1n_2)^{1/4}$ is necessary in the radius of the confidence ball to provide coverage uniformly over the whole space $\mathbb{R}^{\boldsymbol{n}}$.

To handle some technical issues, we impose the following additional condition.

CONDITION (C3). Besides X given by (4.1), we also observe $X'\in\mathbb{R}^{\boldsymbol{n}}$ independent of X, where $X'=\theta+\xi'$, the random vector ξ' satisfies the following relations:

$$\begin{aligned}&\mathbb{P}\big(|\langle v,\xi'\rangle|\ge\sqrt{M}\big)\le\psi_1(M)\quad\text{for all } v\in\mathbb{R}^{\boldsymbol{n}}:\|v\|=1;\\&\mathbb{P}\big(\big|\|\xi'\|^2-V(X')\big|\ge M\sqrt{n_1n_2}\big)\le\psi_2(M),\quad\text{for some statistic } V(X').\end{aligned}\tag{C3}$$

Here $\psi_1(M), \psi_2(M)$ are some positive monotonically decreasing functions such that $\psi_1(M) \downarrow 0$ and $\psi_2(M) \downarrow 0$ as $M \uparrow \infty$.

Typically, $\mathbb{E}\xi'_{ij} = 0$, $\mathrm{Var}(\xi'_{ij}) = 1$, $(i,j) \in [\boldsymbol{n}]$, then $V(X') = |\boldsymbol{n}| = n_1 n_2$. Condition (C3) is satisfied for independent normals $\xi_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$ even if we do not have the sample X' at our disposal. Indeed, in this case we can "duplicate" the original observations X from model (4.1) by randomization at the cost of doubling the variance in the following manner: create samples $X'' = X - Z$ and $X' = X + Z$, for a $Z = (Z_{ij}, (i,j) \in [\boldsymbol{n}])$ (independent of X) such that $Z_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. Relations (C3) are then fulfilled with exponential functions $\psi_l(M) = C_l e^{-c_l M}$ for some $C_l, c_l > 0$, $l = 1,2$, and $V(X') = n_1 n_2$.

If the sub-gaussianity condition (4.8) is fulfilled for ξ' (which is equivalent to Condition (C1) in case of independent ξ'_{ij}'s), then $\psi_1(M) = e^{-\rho M}$. By applying Chebyshev's inequality, it is easy to see that the second relation in (C3) is fulfilled with function $\psi_2(M) = cM^{-2}$ and $V(X') = n_1 n_2$ for any zero mean independent ξ'_{ij}'s with $\mathbb{E}\xi'^2_{ij} = 1$ and $\mathbb{E}[\xi'_{ij}]^4 \le C$.

Coming back to the problem of constructing a confidence ball of full coverage uniformly over $\mathbb{R}^{\boldsymbol{n}}$, let $\hat{\theta}$ be defined as before and based on the sample X. We propose to mimic $\|\theta - \hat{\theta}\|^2$ by the de-biased quantity $\|X' - \hat{\theta}\|^2 - V(X')$ plus additional $\sqrt{n_1 n_2}$-order term to control its oscillations, leading us to the following data dependent radius

$$\tilde{R}^2_M = \big(\|X' - \hat{\theta}\|^2 - V(X') + 2G_M\sqrt{n_1 n_2}\big)_+, \text{ where } G_M = \sqrt{M(M+M_1)}, \tag{4.16}$$

$x_+ = \max\{x, 0\}$ and the constant M_1 is from Theorem 4.1. The next theorem establishes the coverage and size properties of the confidence ball $B(\hat{\theta}, \tilde{R}_M)$.

Theorem 4.3. *Let Conditions* (C1), (C2) *and* (C3) *be fulfilled,* $\tilde{R}^2_M$ *and* G_M *be given by* (4.16), *and* $g_M(\theta, n_1 n_2) = M_1 r^2(\theta) + M + 4G_M\sqrt{n_1 n_2}$. *Then for any* $M \ge 0$

$$\sup_{\theta \in \mathbb{R}^{\boldsymbol{n}}} \mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, \tilde{R}_M)\big) \le \psi_1(M/4) + \psi_2(M) + H_1 e^{-m_1 M},$$
$$\sup_{\theta \in \mathbb{R}^{\boldsymbol{n}}} \mathbb{P}_\theta\big(\tilde{R}^2_M \ge g_M(\theta, n_1 n_2)\big) \le \psi_1(M/4) + \psi_2(M) + 2H_1 e^{-m_1 M},$$

where the constants H_1, m_1, M_1 *are defined in Theorem* 4.1.

The proof of this theorem is the same as the proof of Theorem 5.4, with the only difference that Y, Y' and N should be substituted by X, X' and $n_1 n_2$, also Condition (A4) and Theorem 5.1 should be substituted by Condition (C3) and Theorem 4.1, respectively.

As we already discussed in the beginning of this chapter, in certain situations (e.g., classical *signal+noise* type of models with possible sparsity structure on high-dimensional signal), the deceptiveness issue arises which basically means that for any confidence set it is impossible to have the uniform coverage property and optimal (in this case oracle) size property simultaneously, at least one of the two has to be sacrificed. Interestingly, the deceptiveness issue does not occur in the considered biclustering model (at least to the full extent, namely, modulo a "thin" set $\tilde{\Theta}$ of highly structured parameters defined by (4.14)). Let us explain why.

Recall the notation $N_{\boldsymbol{n}} = n_1 n_2$, the total number of observations. By taking large enough M (which is essentially a multiple by the term $N^{1/2}_{\boldsymbol{n}}$) we can ensure the coverage

and size relations uniformly over the entire space $\mathbb{R}^{\boldsymbol{n}}$. We can interpret the results of Theorem 4.3 as the optimality framework (1.14) with $\Theta_0 = \Theta_1 = \mathbb{R}^{\boldsymbol{n}}$ and the effective radial rate $R(\theta) = r(\theta) + N_{\boldsymbol{n}}^{1/4}$ (for now disregarding the constants and the inflating factor M as we consider only the order of the radial rate). Since the both sets $\Theta_0 = \mathbb{R}^{\boldsymbol{n}}$ and $\Theta_1 = \mathbb{R}^{\boldsymbol{n}}$ are the biggest possible, the only ingredient of the optimality frameworks (1.14) that can be affected by the deceptiveness phenomenon is the effective radial rate $R(\theta)$.

By (4.15), we have that for $\theta \in \mathbb{R}^{\boldsymbol{n}} \backslash \tilde{\Theta}$, the effective squared radial rate is of the oracle rate order: $R(\theta) = N_{\boldsymbol{n}}^{1/4} + r(\theta) \le C r(\theta)$. So, the effective radial rate $R(\theta)$ can be of a worse order than the oracle rate $r(\theta)$ only for highly structured parameters $\theta \in \tilde{\Theta}$ given by (4.14). This minor defect of the radial rate $R(\theta)$ is the only sacrifice in the size relation of the optimality framework (1.14) for the biclustering model. We can also see this as the optimality framework (1.14) with $\Theta_0 = \mathbb{R}^{\boldsymbol{n}}$, $\Theta_1 = \mathbb{R}^{\boldsymbol{n}} \backslash \tilde{\Theta}$ and the effective radial rate $r(\theta)$, then a minor sacrifice occurs in the set $\Theta_1 = \mathbb{R}^{\boldsymbol{n}} \backslash \tilde{\Theta}$ which is slightly "smaller" than the entire space $\mathbb{R}^{\boldsymbol{n}}$.

Excluding the whole set $\tilde{\Theta}$ is actually too precautios, not all θ from $\tilde{\Theta}$ are problematic. Only those $\theta \in \tilde{\Theta}$ cause troubles for which $r(\theta)$ is of a smaller order than $N_{\boldsymbol{n}}^{1/4}$ (in a certain high-dimensional asymptotic setting). Precisely, this happens if $k_1(\theta) = 1$ and $n_2 \ll n_1$ (one row whose dimension dominates the number of columns), or, if $k_2(\theta) = 1$ and $n_1 \ll n_2$. In the high-dimensional setting, $n_1 \ll n_2$ means $n_2(l)/n_1(l) \to \infty$ as $l \to \infty$ when $n_1 = n_1(l) \to \infty$ and $n_2 = n_2(l) \to \infty$ as $l \to \infty$. Essentially, the first case corresponds to one lengthy row, the second case to one lengthy column, these are indeed "highly structured" parameters. We conjecture that it is in general impossible to construct a confidence set satisfying the optimality framework (1.14) with $\Theta_0 = \Theta_1 = \mathbb{R}^{\boldsymbol{n}}$ and the effective local radial rate equal to the oracle rate $r(\theta)$, a small portion of "highly structured" parameters must be removed.

Remark 4.10. *In way, one can say that the biclustering model is "too difficult" (or too uninformative) for the deceptiveness phenomenon to (fully) occur in this model. The oracle rate turns out to be too big for the biclustering model to suffer from the deceptiveness issue.*

4.3. IMPLICATIONS

4.3.1. MINIMAX RESULTS FOR THE BICLUSTERING MODEL

The above local results imply adaptive (global) minimax results for estimation and uncertainty quantification problems over all biclustering and graphon scales (of classes) $\Theta_{\mathcal{S}} = \{\Theta_s, s \in \mathcal{S}\}$ whose minimax rate $r^2(\Theta_s) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta_s} \mathbb{E}_\theta \|\tilde{\theta} - \theta\|^2$ is bounded from below by a multiple of the oracle rate: $r^2(\Theta_s) \ge c r^2(\theta)$ for all $\theta \in \Theta_s$, $s \in \mathcal{S}$; we say that $r(\theta)$ *covers* the scale $\Theta_{\mathcal{S}}$. If this holds, then the adaptive (with respect to the structural parameter $s \in \mathcal{S}$) minimax estimation result follows immediately from Theorem 4.1:

$$\sup_{\theta \in \Theta_s} \mathbb{P}_\theta \big(\|\hat{\theta} - \theta\|^2 \ge \tfrac{M_1}{c} r^2(\Theta_s) + M \big) \le H_1 e^{-m_1 M}.$$

The minimax results for the uncertainty quantification problem follow by the same argument.

In particular, we can derive the same type of adaptive minimax estimation results as in [41], as a consequence of our local results. In [41] (cf. [40]), classes $\Theta^{\text{asym}}_{k_1 k_2}$ are intro-

duced. In our notation, $\Theta^{\rm asym}_{k_1k_2} = \Theta(\boldsymbol{k}) \triangleq \cup_{\boldsymbol{I}\in\mathcal{I}_{\boldsymbol{k}}}\Theta_{\boldsymbol{I}}$, where $\boldsymbol{k} = (k_1,k_2)$, $\Theta_{\boldsymbol{I}} \triangleq \mathbb{L}_{\boldsymbol{I}} \cap [0,1]^{\boldsymbol{n}}$ and $\mathbb{L}_{\boldsymbol{I}}$ is defined by (4.2). So, the family of classes $\Theta^{\rm asym}_{k_1k_2}$ is nothing else but the scale $\{\Theta(\boldsymbol{k}), \boldsymbol{k}\in[\boldsymbol{n}]\}$. The minimax rate $r^2(\Theta(\boldsymbol{k})) = k_1k_2 + n_1\log k_1 + n_2\log k_2$ over $\Theta(\boldsymbol{k})$ is derived in [40], under the assumption $\log k_1 \asymp \log k_2$. In our notation: there exist $c>0$ and $\delta\in(0,1]$ such that

$$\inf_{\tilde\theta}\sup_{\theta\in\Theta(\boldsymbol{k})} \mathbb{P}_\theta\big(\|\tilde\theta-\theta\|^2 \ge c r^2(\Theta(\boldsymbol{k}))\big) \ge \delta.$$

It is easy to see that for each $\boldsymbol{k}\in[\boldsymbol{n}]$ the minimax rate over $\Theta(\boldsymbol{k})$ is bigger than the oracle rate $r^2(\theta)$ for any $\theta\in\Theta(\boldsymbol{k})$. Indeed, if $\theta\in\Theta(\boldsymbol{k})$, then $\theta\in\mathbb{L}_{\boldsymbol{I}}$ for some $\boldsymbol{I}\in\mathcal{I}_{\boldsymbol{k}}$, so that $\mathrm{P}_{\boldsymbol{I}}\theta=\theta$ and hence

$$r^2(\theta) \le r^2(\boldsymbol{I},\theta) = \rho(\boldsymbol{k}(\boldsymbol{I})) = \rho(\boldsymbol{k}) \le r^2(\Theta(\boldsymbol{k})), \quad \theta\in\Theta(\boldsymbol{k}), \tag{4.17}$$

where the quantity $\rho(\boldsymbol{k})$ is defined by (4.5). Then Theorems 4.1 and 4.3 immediately imply the adaptive (for this scale) minimax estimation result and the size relation in the uncertainty quantification problem, which is summarized by the following corollary.

Corollary 4.1. *Let the conditions of Theorems 4.1 and 4.3 be fulfilled, $g_M(\Theta(\boldsymbol{k})) = M_1 r^2(\Theta(\boldsymbol{k})) + M + 4G_M\sqrt{n_1n_2}$. Then for any $\boldsymbol{k}\in[\boldsymbol{n}]$ and $M\ge 1$,*

$$\sup_{\theta\in\Theta(\boldsymbol{k})} \mathbb{P}_\theta\big(\|\hat\theta-\theta\|^2 \ge M_1 r^2(\Theta(\boldsymbol{k})) + M\big) \le H_1 e^{-m_1M},$$
$$\sup_{\theta\in\Theta(\boldsymbol{k})} \mathbb{P}_\theta\big(\tilde R^2_M \ge g_M(\Theta(\boldsymbol{k}))\big) \le \psi_1(M/4) + \psi_2(M) + 2H_1e^{-m_1M}.$$

The first claim of the corollary recovers (in a sligtly more general formulation) the same minimax estimation rate $r^2(\Theta(\boldsymbol{k}))$ as in [41]. We do not specialize Theorem 4.2 and the coverage relation of Theorem 4.3 for the scale $\{\Theta(\boldsymbol{k}), \boldsymbol{k}\in[\boldsymbol{n}]\}$ because both Theorem 4.2 and the coverage relation of Theorem 4.3 hold uniformly in $\theta\in\mathbb{R}^{\boldsymbol{n}}$, hence uniformly over any $\Theta(\boldsymbol{k})$.

Remark 4.11. *From (4.17), we have that $r^2(\theta) \le k_1k_2 + n_1\log k_1 + n_2\log k_2$ for each $\theta\in\Theta(\boldsymbol{k})$. But for any $\boldsymbol{I}\in\mathcal{I}_{\boldsymbol{k}}$ with $\boldsymbol{k}=(k_1,k_2)$ and any $\mathcal{L}_{\boldsymbol{I}}$, there exist $\boldsymbol{I}'=\boldsymbol{I}'(\boldsymbol{I})$ and $\mathcal{L}_{\boldsymbol{I}'}$ such that $\mathcal{L}_{\boldsymbol{I}}\subseteq\mathcal{L}_{\boldsymbol{I}'}$ where $\boldsymbol{I}'\in\mathcal{I}_{\boldsymbol{k}'}$ with $\boldsymbol{k}'=(k_1,n_2)$. Precisely, if $\boldsymbol{I}=\cup_{(i,j)\in[\boldsymbol{k}]}(I_i,J_j)$, take $\boldsymbol{I}'=\cup_{(i,j)\in[\boldsymbol{k}']}(I_i,\{j\})$. So, for any $\theta\in\Theta(\boldsymbol{k})$, $\theta\in\mathbb{L}_{\boldsymbol{I}}\subseteq\mathcal{L}_{\boldsymbol{I}'(\boldsymbol{I})}$ for some $\boldsymbol{I}'\in\mathcal{I}_{\boldsymbol{k}'}$ with $\boldsymbol{k}'=(k_1,n_2)$, implying $\mathrm{P}_{\boldsymbol{I}'}\theta=\theta$ and (in view of (4.5)) $\rho(\boldsymbol{k}')=k_1n_2+n_1\log k_1$. Then by the oracle definition, $r^2(\theta)\le r^2(\boldsymbol{I}',\theta)=\rho(\boldsymbol{k}(\boldsymbol{I}'))=\rho(\boldsymbol{k}')=k_1n_2+n_1\log k_1$ for all $\theta\in\Theta(\boldsymbol{k})$. Similarly, we can derive that $r^2(\theta)\le n_1k_2+n_2\log k_2$ and $r^2(\theta)\le n_1n_2$ for all $\theta\in\Theta(\boldsymbol{k})$.*

Thus, instead of (4.17), we actually established the following stronger bound:

$$r^2(\theta) \le \bar r^2(\Theta(\boldsymbol{k})) \triangleq \min\{k_1k_2+n_1\log k_1+n_2\log k_2, k_1n_2+n_1\log k_1, n_1k_2+n_2\log k_2, n_1n_2\}.$$

for all $\theta\in\Theta(\boldsymbol{k})$. Notice that, for some $\boldsymbol{k}\in[\boldsymbol{n}]$, $\bar r^2(\Theta(\boldsymbol{k}))$ can be less than the minimax rate $r^2(\Theta(\boldsymbol{k})) = k_1k_2+n_1\log k_1+n_2\log k_2$ which is claimed in [40] only under the assumption $\log k_1\asymp\log k_2$. Indeed, in this case $r^2(\Theta(\boldsymbol{k}))\asymp\bar r^2(\Theta(\boldsymbol{k}))$. Clearly, the minimax rate over $\Theta(\boldsymbol{k})$ for arbitrary $\boldsymbol{k}\in[\boldsymbol{n}]$ cannot be bigger than $\bar r^2(\Theta(\boldsymbol{k}))$, we conjecture that it is $\bar r^2(\Theta(\boldsymbol{k}))$ for all $\boldsymbol{k}\in[\boldsymbol{n}]$.

Remark 4.12. *We can formulate the minimax results of the above corollary in terms of the "finer" scale* $\{\Theta_{\boldsymbol{I}}, \boldsymbol{I} \in \mathcal{I}\}$, *with* $\Theta_{\boldsymbol{I}} \triangleq \mathbb{L}_{\boldsymbol{I}} \cap [0,1]^{\boldsymbol{n}}$.

Remark 4.13. *As to the deceptiveness phenomenon, the same situation occurs for the minimax adaptive version of uncertainty quantification problem over the scale* $\{\Theta(\boldsymbol{k}), \boldsymbol{k} \in [\boldsymbol{n}]\}$ *as for the local version discussed in Section 4.2.2. Namely, we can claim the coverage and the size relations uniformly over the whole scale* $\{\Theta(\boldsymbol{k}), \boldsymbol{k} \in [\boldsymbol{n}]\}$, *but the effective radial rate* $R(\theta) = \sqrt{g_M(\Theta(\boldsymbol{k}))}$ *in the size relation is* $r(\Theta(\boldsymbol{k})) + N_{\boldsymbol{n}}^{1/4}$, *not* $r(\Theta(\boldsymbol{k}))$. *Certainly,* $g_M(\Theta(\boldsymbol{k})) \asymp r^2(\Theta(\boldsymbol{k}))$ *for the subscale* $\{\Theta(\boldsymbol{k}), \boldsymbol{k} \in [\boldsymbol{n}], \min(k_1, k_2) \neq 1\}$, *with highly structured classes removed (*$\Theta(\boldsymbol{k})$ *with* $k_1 = 1$ *or* $k_2 = 1$*). Actually only for some of the highly structured classes there is a sacrifice in the radial rate. Precisely, the effective radial rate is of a bigger order than the minimax rate:* $R(\theta) = \sqrt{g_M(\Theta(\boldsymbol{k}))} \gg r(\Theta(\boldsymbol{k}))$, *only for* $\Theta(\boldsymbol{k})$, *with* $k_1 = 1$ *and* $n_2 \ll n_1$, *or with* $k_2 = 1$ *and* $n_1 \ll n_2$.

Remark 4.14. *Although the main focus of this chapter is the uncertainty quantification problem, let us relate our estimation results to the results of the paper [41]. In [41], partially observed data are allowed, our results can also be extended to this more general setting by an appropriate adjustment of Condition* (C1). *Next, we derived our estimation results by using the penalization method with appropriately chosen penalty whereas the method in [41] is based on the constrained least squares estimation and the 2-fold cross validation technique; our exponential probability bound formulation is slightly more refined. As is already mentioned in the beginning of this chapter, another distinctive feature of our results is that they are local and imply the same type of global minimax results as in [41]. Finally, in [41] the coordinates of the vector* ξ *must be independent and satisfy the sub-gaussianity condition, while we only assume weaker Condition* (C1), *which allows dependent coordinates of* ξ*; for instance, the* ξ_i*'s can originate from an autoregressive model.*

4.3.2. STOCHASTIC BLOCK MODEL (SBM)

Here we briefly discuss a particular case of biclustering model, the *stochastic block model* (SBM) which is used in the literature on networks to model undirected network graphs. Precisely, to get the SBM from the biclustering model, we set $X_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$, $k_1 = k_2 = k$, $n_1 = n_2 = n$, $z_1 = z_2 = z$. For a mapping $z \in [k]^{[n]}$, the pertinent partition in the SBM is $\boldsymbol{I} = \boldsymbol{I}(z) = \cup_{i,j \in [k]}(z^{-1}(i), z^{-1}(j))$, which is in turn used in (4.2). Since $X_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$ for $i > j$ stands for the presence or absence of an edge between vertices i and j in the network interpretation, to model undirected network graphs, some conditions (called *network conditions*) are additionally assumed: the "no self-loop" condition $X_{ii} = \theta_{ii} = 0$ and the symmetry condition $X_{ij} = X_{ji}$ and $\theta_{ij} = \theta_{ji}$. Denote by Θ_{net} the parameters $\theta \in \mathbb{R}^{\boldsymbol{n}}$ satisfying these additional network conditions.

All the quantities and claims specialize to the SBM by setting $k_1 = k_2 = k$, $n_1 = n_2 = n$, $z_1 = z_2 = z$ in all the above formulas for the biclustering model. The linear subspaces $\mathbb{L}_{\boldsymbol{I}}$ defined by (4.2) will get adjusted since $z_1 = z_2$, the complexity layer $\mathcal{I}_{\boldsymbol{k}}$ is the collection of all possible partitions of $([n], [n])$ into (k, k) blocks, parametrized by mappings $z \in [k]^{[n]}$. Clearly, $|\mathcal{I}_k| \leq |\bar{\mathcal{I}}_k| = k^n$, $k \in [n]$, and, for each $\boldsymbol{I} \in \mathcal{I}_k$, $\dim(\mathbb{L}_{\boldsymbol{I}}) = k^2$. Notice that under additional network conditions $ck^2 \leq \dim(\mathbb{L}_{\boldsymbol{I}}) = (k-1)k/2 \leq k^2$ for each $\boldsymbol{I} \in \mathcal{I}_k$, so that we use k^2 (instead of true $\dim(\mathbb{L}_{\boldsymbol{I}})$) in the complexity part of the local rate as it is still of the

same order, although some constants could have been improved because of this extra network structure.

Denote by $k(\boldsymbol{I})$ the number of nonempty row (and column) blocks in partition $\boldsymbol{I}$. The complexity of $\boldsymbol{I} \in \mathcal{I}_k$ becomes $\rho(k(\boldsymbol{I})) = \rho(k) = k^2 + n\log k$. The corresponding family of *local rates* becomes

$$r^2(\boldsymbol{I},\theta) = \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \rho(k(\boldsymbol{I})), \quad \boldsymbol{I} \in \mathcal{I}.$$

The oracle $(\boldsymbol{I}_o, k_o) = (\boldsymbol{I}_o(\theta), k_o(\theta))$ (where $k_o = k(\boldsymbol{I}_o)$) is defined by

$$r^2(\theta) = \min_{\boldsymbol{I}\in\mathcal{I}} r^2(\boldsymbol{I},\theta) = \|\theta - \mathrm{P}_{\boldsymbol{I}_o}\theta\|^2 + \rho(k_o). \tag{4.18}$$

Theorems 4.1, 4.2 and 4.3 immediately imply the local (oracle) results for SBM (estimation, structure recovery, the coverage and size properties), which are summarized in the following corollary. We keep the same notation for all the quantities and constants involved as for the biclustering model, with the understanding that these are specialized for the SBM and some constants must be adjusted.

Corollary 4.2. *Let the conditions of Theorems 4.1 and 4.3 be fulfilled, $g_M(\theta,n) = M_1 r^2(\theta) + M + 4G_M n$. Then for any $M \ge 1$,*

$$\begin{aligned}
&\sup_{\theta\in\mathbb{R}^{n^2}} \mathbb{P}_\theta\big(\|\hat\theta - \theta\|^2 \ge M_1 r^2(\theta) + M\big) \le H_1 e^{-m_1 M},\\
&\sup_{\theta\in\mathbb{R}^{n^2}} \mathbb{P}_\theta\big(r^2(\hat{\boldsymbol{I}},\theta) \ge c_3 r^2(\theta) + M\big) \le \Sigma(c_1) e^{-c_2 M},\\
&\sup_{\theta\in\mathbb{R}^{n^2}} \mathbb{P}_\theta\big(\theta \notin B(\hat\theta, \tilde R_M)\big) \le \psi_1(M/4) + \psi_2(M) + H_1 e^{-m_1 M},\\
&\sup_{\theta\in\mathbb{R}^{n^2}} \mathbb{P}_\theta\big(\tilde R_M^2 \ge g_M(\theta,n)\big) \le \psi_1(M/4) + \psi_2(M) + 2H_1 e^{-m_1 M}.
\end{aligned}$$

Now consider some adaptive minimax results for the SBM which follow from Corollary 4.2. In [40] (cf. [54]), classes Θ_k were introduced. In our notation, $\Theta_k = \cup_{\boldsymbol{I}\in\mathcal{I}_k}\Theta_{\boldsymbol{I}}$, where $\Theta_{\boldsymbol{I}} = \mathbb{L}_{\boldsymbol{I}} \cap \Theta_{\mathrm{net}} \cap [0,1]^{n^2}$, $\boldsymbol{I} \in \mathcal{I}_k$, $k \in [n]$. So, we have the scale $\{\Theta_k, k \in [n]\}$ and the adaptive minimax results over this scale follow from the local results given by Corollary 4.2. Indeed, as is shown in [40], the minimax rate over Θ_k in the SBM is

$$\inf_{\tilde\theta} \sup_{\theta\in\Theta_k} \mathbb{E}_\theta\|\tilde\theta - \theta\|^2 \asymp k^2 + n\log k \triangleq r^2(\Theta_k).$$

On the other hand, for each $\theta \in \Theta_k$ there exists $\boldsymbol{I} \in \mathcal{I}_k$ such that $\theta \in \mathbb{L}_{\boldsymbol{I}}$. Hence, $\mathrm{P}_{\boldsymbol{I}}\theta = \theta$ and $r^2(\theta) \le r^2(\boldsymbol{I},\theta) = \rho(k) = r^2(\Theta_k)$. We obtain the following corollary.

Corollary 4.3. *Let the conditions of Theorems 4.1 and 4.3 be fulfilled, and $g_M(\Theta_k) = M_1 r^2(\Theta_k) + M + 4G_M n$. Then for any $k \in [n]$ and $M \ge 1$,*

$$\begin{aligned}
&\sup_{\theta\in\Theta_k} \mathbb{P}_\theta\big(\|\hat\theta - \theta\|^2 \ge M_1 r^2(\Theta_k) + M\big) \le H_1 e^{-m_1 M},\\
&\sup_{\theta\in\Theta_k} \mathbb{P}_\theta\big(\tilde R_M^2 \ge g_M(\Theta_k)\big) \le \psi_1(M/4) + \psi_2(M) + 2H_1 e^{-m_1 M}.
\end{aligned}$$

The first claim recovers (in a slightly more general formulation) the same minimax estimation rate $r^2(\Theta_k)$ as in [40] and [54]. We do not specialize Theorem 4.2 and the coverage relation of Theorem 4.3 for the scale $\{\Theta_k, k \in [n]\}$ because it does not make sense to specialize these claims for any scale.

Remark 4.15. *We can formulate the minimax results for the finer scale* $\{\Theta_{\boldsymbol{I}}, \boldsymbol{I} \in \mathcal{I}\}$, $\Theta_{\boldsymbol{I}} = \mathbb{L}_{\boldsymbol{I}} \cap \Theta_{\text{net}} \cap [0,1]^{n^2}$.

Remark 4.16. *As to the deceptiveness phenomenon in the SBM, we again have the coverage property uniformly over the whole scale* $\{\Theta_k, k \in [n]\}$, *whereas the size property with the optimal radial rate holds over all classes* $\{\Theta_k, k = 2, \dots, n\}$, *but one:* Θ_1. *The class* Θ_1 *indeed consists of highly structured parameters* $\theta \in \mathbb{R}^{n^2}$, *whose all coordinates are equal. This is just a one-dimensional signal+noise model with* n^2 *observations. Clearly,* $g_M(\Theta_1) \asymp n \gg 1 = r^2(\Theta_1)$, *it seems impossible to mimic the optimal rate in the one-dimensional model.*

Finally, we derive the global minimax results for the function class of graphons as consequence of our local results, Theorems 4.1, 4.2 and 4.3. We use the same notation as in [40]. Consider a random graph with adjacency matrix $\{X_{ij}\} \in \{0,1\}^{\boldsymbol{n}}$. Assume again the network conditions: $X_{ii} = \theta_{ii} = 0$, $X_{ij} = X_{ji}$, $\theta_{ij} = \theta_{ji}$. For any $i > j$, X_{ij} is sampled as follows:

$$(\xi_1, \dots, \xi_n) \sim \mathrm{P}_\xi, \quad X_{ij} | (\xi_i, \xi_j) \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = f(\xi_i, \xi_j).$$

The function f on $[0,1]^2$, which is assumed to be symmetric, is called *graphon*. Introduce the derivative operator $\nabla_{jk} f(x,y) = \frac{\partial^{j+k}}{\partial x^j \partial y^k} f(x,y)$, with the convention $\nabla_{00} f(x,y) = f(x,y)$. For $\alpha > 0$, the Hölder norm is defined by

$$\|f\|_{\mathcal{H}_\alpha} = \max_{j+k \le \lfloor \alpha \rfloor} \sup_{x,y} |\nabla_{jk} f(x,y)| + \max_{j+k = \lfloor \alpha \rfloor} \sup_{(x,y) \ne (x',y')} \frac{|\nabla_{jk} f(x,y) - \nabla_{jk} f(x',y')|}{(|x-x'| + |y-y'|)^{\alpha - \lfloor \alpha \rfloor}}.$$

For $\alpha, Q > 0$, the Hölder graphon class is

$$\mathcal{F}_\alpha(Q) = \{f : \|f\|_{\mathcal{H}_\alpha} \le Q,\ f(x,y) = f(y,x),\ 0 \le f(x,y) \le 1 \text{ for } x \ge y\}.$$

Recall that $\theta_{ij} = f(\xi_i, \xi_j)$. Slightly abusing notation, we will write $\theta \in \mathcal{F}_\alpha(Q)$ if $f \in \mathcal{F}_\alpha(Q)$.

The next proposition is Lemma 2.1 from [40], which we give here (in our notation) for completeness. The proof can be found in [40].

Proposition 4.1. *For any* $\theta \in \mathcal{F}_\alpha(Q)$, $k_0 \in [n]$, *there exists a partition* $\boldsymbol{I}_0 = \boldsymbol{I}_0(\theta, k_0) \in \mathcal{I}_{k_0}$ *such that, for some universal constant* $\bar{C}_1 > 0$,

$$\|\theta - \mathrm{P}_{\boldsymbol{I}_0} \theta\|^2 \le \bar{C}_1 Q^2 n^2 k_0^{-2\min\{\alpha,1\}}.$$

By taking $k_0 = \lceil n^{1/(\min\{\alpha,1\}+1)} \rceil$ and $\boldsymbol{I}_0 \in \mathcal{I}_{k_0}$ from Proposition 4.1, we obtain for some $\bar{C}_2 > 0$ that

$$\begin{aligned}
\sup_{\theta \in \mathcal{F}_\alpha(Q)} r^2(\theta) &= \sup_{\theta \in \mathcal{F}_\alpha(Q)} \left\{ \|\theta - P_{\boldsymbol{I}_o} \theta\|^2 + k_o^2(\theta) + n \log k_o(\theta) \right\} \\
&\le \sup_{\theta \in \mathcal{F}_\alpha(Q)} \|\theta - P_{\boldsymbol{I}_0} \theta\|^2 + k_0^2 + n \log k_0 \le \bar{C}_1 Q^2 n^2 k_0^{-2\min\{\alpha,1\}} + k_0^2 + n \log k_0 \\
&\le \bar{C}_2 \left(n^{2-2\alpha/(\alpha+1)} + n \log n \right) = \bar{C}_2 \left(n^{2/(\alpha+1)} + n \log n \right). \qquad (4.19)
\end{aligned}$$

Then Corollary 4.2 implies the following.

Corollary 4.4. *Let the conditions of Theorems 4.1 and 4.3 be fulfilled,* $\bar{M}_1 = M_1 \bar{C}_2^{-1}$ *(M_1 is from Corollary 4.2) and $\bar{C}_2$ be from (4.19). Then for any $M \geq 1$,*

$$\sup_{\theta \in \mathcal{F}_\alpha(Q)} \mathbb{P}_\theta\bigl(\|\hat{\theta} - \theta\|^2 \geq \bar{M}_1 (n^{2/(\alpha+1)} + n \log n) + M\bigr) \leq H_1 e^{-m_1 M},$$

$$\sup_{\theta \in \mathcal{F}_\alpha(Q)} \mathbb{P}_\theta\bigl(\tilde{R}_M^2 \geq \bar{M}_1 (n^{2/(\alpha+1)} + n \log n) + M + 4G_M n\bigr) \leq \psi_1(\tfrac{M}{4}) + \psi_2(M) + 2H_1 e^{-m_1 M}.$$

The first claim recovers (in a more general formulation) the same minimax estimation rate as in [40] and [54]. We again do not specialize Theorem 4.2 and the coverage relation of Theorem 4.3, these claims remain the same as in Corollary 4.2.

4

4.4. TECHNICAL LEMMAS

We provide a couple of technical lemmas used in the proofs of the main results.

Lemma 4.1. *Let Condition (C1) be fulfilled and the structure selector $\hat{\boldsymbol{I}}$ is defined by (4.9). Then for any $\theta \in \mathbb{R}^{\boldsymbol{n}}$ and any $\boldsymbol{I}, \boldsymbol{I}_0 \in \mathcal{I}$,*

$$\mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) \leq \exp\Bigl\{ - A\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + B\|\theta - \mathrm{P}_{\boldsymbol{I}_0}\theta\|^2 - C\rho(\boldsymbol{k}(\boldsymbol{I})) + D\rho(\boldsymbol{k}(\boldsymbol{I}_0))\Bigr\},$$

where the constants $A = \frac{\alpha}{16}$, $B = \frac{3\alpha}{16}$ and $C = \frac{K\alpha - 5}{8}$, $D = \frac{3 + K\alpha}{8}$.

Moreover, if $\mathbb{L}_{\boldsymbol{I}} \subseteq \mathbb{L}_{\boldsymbol{I}_0}$, then

$$\mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) \leq \exp\Bigl\{ - \tfrac{\alpha}{3}\theta^T(\mathrm{P}_{\boldsymbol{I}_0} - \mathrm{P}_{\boldsymbol{I}})\theta - \tfrac{K\alpha}{2}\rho(\boldsymbol{k}(\boldsymbol{I})) + (1 + \tfrac{K\alpha}{2})\rho(\boldsymbol{k}(\boldsymbol{I}_0))\Bigr\}.$$

Proof. Recall that $\mathrm{P}_{\boldsymbol{I}}$ is the projection onto $\mathbb{L}_{\boldsymbol{I}}$. For any $\boldsymbol{I}, \boldsymbol{I}_0 \in \mathcal{I}$, we have that

$$\begin{aligned} X^T(\mathrm{P}_I - \mathrm{P}_{I_0})X &= \theta^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\theta + 2\theta^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\xi + \xi^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\xi \\ &\leq -\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \|\theta - \mathrm{P}_{\boldsymbol{I}_0}\theta\|^2 + 2|\theta^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\xi| + \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 - \|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2. \end{aligned} \tag{4.20}$$

Using the relations $\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0} = (\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}} + \mathbb{L}_{\boldsymbol{I}_0}}$, $\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}} + \mathbb{L}_{\boldsymbol{I}_0}} x\|^2 \leq \|\mathrm{P}_{\boldsymbol{I}} x\|^2 + \|\mathrm{P}_{\boldsymbol{I}_0} x\|^2$, $x \in \mathbb{R}^{\boldsymbol{n}}$, and the inequality $2ab \leq a^2/4 + 4b^2$ (for any $a, b \in \mathbb{R}$), we derive

$$\begin{aligned} &2|\theta^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\xi| = 2|\theta^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}} + \mathbb{L}_{\boldsymbol{I}_0}}\xi| \leq 2\|\theta^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\|\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}} + \mathbb{L}_{\boldsymbol{I}_0}}\xi\| \\ &\leq \tfrac{1}{4}\|(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})\theta\|^2 + 4\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}} + \mathbb{L}_{\boldsymbol{I}_0}}\xi\|^2 = \tfrac{1}{4}\|\theta - \mathrm{P}_{\boldsymbol{I}_0}\theta - (\theta - \mathrm{P}_{\boldsymbol{I}}\theta)\|^2 + 4\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}} + \mathbb{L}_{\boldsymbol{I}_0}}\xi\|^2 \\ &\leq \tfrac{1}{2}\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \tfrac{1}{2}\|\theta - \mathrm{P}_{\boldsymbol{I}_0}\theta\|^2 + 4\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 + 4\|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2. \end{aligned}$$

The last bound and (4.20) imply that

$$X^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})X \leq -\tfrac{1}{2}\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \tfrac{3}{2}\|\theta - \mathrm{P}_{\boldsymbol{I}_0}\theta\|^2 + 5\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 + 3\|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2. \tag{4.21}$$

Combining (4.9) and (4.21) with the Markov inequality, we derive for any $\boldsymbol{I}, \boldsymbol{I}_0 \in \mathcal{I}$, $h \geq 0$,

$$\begin{aligned}
&\mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) \\
&\leq \mathbb{P}_\theta\big(\exp\{\|X - \mathrm{P}_{\boldsymbol{I}_0} X\|^2 + K\rho(\boldsymbol{k}(\boldsymbol{I}_0))\} - \|X - \mathrm{P}_{\boldsymbol{I}} X\|^2 - K\rho(\boldsymbol{k}(\boldsymbol{I}))\big\} \geq 1\big) \\
&\leq \mathbb{E}_\theta \exp\big\{h\big(X^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})X + K[\rho(\boldsymbol{k}(\boldsymbol{I}_0)) - \rho(\boldsymbol{k}(\boldsymbol{I}))]\big)\big\} \qquad (4.22) \\
&\leq \exp\big\{-\tfrac{h}{2}\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \tfrac{3h}{2}\|\theta - \mathrm{P}_{\boldsymbol{I}_0}\theta\|^2 + hK[\rho(\boldsymbol{k}(\boldsymbol{I}_0)) - \rho(\boldsymbol{k}(\boldsymbol{I}))]\big\} \\
&\quad \times \mathbb{E}_\theta \exp\big\{5h\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 + 3h\|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2\big\}.
\end{aligned}$$

The first claim of the lemma follows for $h = \frac{\alpha}{8}$ from the last display and the relation

$$\begin{aligned}
\mathbb{E}_\theta \exp\big\{\tfrac{5\alpha}{8}\|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 + \tfrac{3\alpha}{8}\|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2\big\} &\leq \big[\mathbb{E}_\theta \exp\{\alpha|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\}\big]^{5/8}\big[\mathbb{E}_\theta \exp\{\alpha|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2\}\big]^{3/8} \\
&\leq \exp\big\{\tfrac{5}{8}\rho(\boldsymbol{k}(\boldsymbol{I})) + \tfrac{3}{8}\rho(\boldsymbol{k}(\boldsymbol{I}_0))\big\},
\end{aligned}$$

which is in turn obtained by using the Hölder inequality and Condition (C1).

In the case $\mathbb{L}_{\boldsymbol{I}} \subseteq \mathbb{L}_{\boldsymbol{I}_0}$, as $(a+b)^2 \geq 2a^2/3 - 2b^2$ and $\mathrm{P}_{\boldsymbol{I}_0} - \mathrm{P}_{\boldsymbol{I}} = \mathrm{P}_{\mathbb{L}_{\boldsymbol{I}}^\perp \cap \mathbb{L}_{\boldsymbol{I}_0}}$, we obtain

$$\begin{aligned}
X^T(\mathrm{P}_{\boldsymbol{I}} - \mathrm{P}_{\boldsymbol{I}_0})X = -\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}}^\perp \cap \mathbb{L}_{\boldsymbol{I}_0}} X\|^2 &\leq -\tfrac{2}{3}\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}}^\perp \cap \mathbb{L}_{\boldsymbol{I}_0}}\theta\|^2 + 2\|\mathrm{P}_{\mathbb{L}_{\boldsymbol{I}}^\perp \cap \mathbb{L}_{\boldsymbol{I}_0}}\xi\|^2 \\
&\leq \tfrac{2}{3}\big(\|\mathrm{P}_{\boldsymbol{I}}\theta\|^2 - \|\mathrm{P}_{\boldsymbol{I}_0}\theta\|^2\big) + 2\|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2 = -\tfrac{2}{3}\theta^T(\mathrm{P}_{\boldsymbol{I}_0} - \mathrm{P}_{\boldsymbol{I}})\theta + 2\|\mathrm{P}_{\boldsymbol{I}_0}\xi\|^2.
\end{aligned}$$

We use the last relation and Condition (C1) in (4.22) with $h = \alpha/2$ to derive the second claim of the lemma. □

Note that above lemma holds for any $\boldsymbol{I}_0 \in \mathcal{I}$. By taking $\boldsymbol{I}_0 = \boldsymbol{I}_o$ defined by (4.12), we obtain the following lemma.

Lemma 4.2. *Let Conditions* (C1) *and* (C2) *be fulfilled. Then there exist positive constants* $c_1 = c_1(K) \geq 2$, c_2 *and* $c_3 = c_3(K)$ *such that for any* $\theta \in \mathbb{R}^{\boldsymbol{n}}$

$$\mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) \leq \exp\big\{-c_1\rho(\boldsymbol{k}(\boldsymbol{I})) - c_2[r^2(\boldsymbol{I},\theta) - c_3 r^2(\theta)]\big\}.$$

Proof. With constants A, B, C, D defined in Lemma 4.1, define the constant $c_1 = c_1(K) = C - A = \frac{K\alpha - 5}{8} - \frac{\alpha}{16} \geq 2$ since $K \geq \bar{K} = \frac{1}{2} + \frac{21}{\alpha}$ by Condition (C2). Applying this and Lemma 4.1 with $\boldsymbol{I}_0 = \boldsymbol{I}_o$, we derive that

$$\begin{aligned}
\mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) &\leq \exp\big\{-c_1\rho(\boldsymbol{k}(\boldsymbol{I})) - \big[Ar^2(\boldsymbol{I},\theta) - \max(B,D)r^2(\theta)\big]\big\} \\
&= \exp\big\{-c_1\rho(\boldsymbol{k}(\boldsymbol{I})) - c_2\big[r^2(\boldsymbol{I},\theta) - c_3 r^2(\theta)\big]\big\},
\end{aligned}$$

which completes the proof with the constants $c_1 = \frac{K\alpha-5}{8} - \frac{\alpha}{16} \geq 2$, $c_2 = A = \frac{\alpha}{16}$ and $c_3 = c_3(K) = A^{-1}\max(B,D) = \frac{16}{\alpha}\max\big(\frac{3\alpha}{16}, \frac{3+K\alpha}{8}\big) = \max\big\{3, \frac{6}{\alpha} + 2K\big\} = \frac{6}{\alpha} + 2K$. □

4.5. PROOFS OF THE THEOREMS

In this section we gather the proofs of all the theorems. By C_1, C_2 etc., we denote constants which can only depend on α and K.

Proof of Theorem 4.1. Recall the constants c_1, c_2, c_3 defined in the proof of Lemma 4.2. Denote for brevity $\Delta_M = \Delta_M(\theta) = M_1 r^2(\theta) + M$, $R_{\boldsymbol{I}}^2 = R_{\boldsymbol{I}}^2(\theta, X) = \|\theta - \mathrm{P}_{\boldsymbol{I}} X\|^2 = \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2$ and $\hat{p}_{\boldsymbol{I}} = 1\{\hat{\boldsymbol{I}} = \boldsymbol{I}\}$, $\boldsymbol{I} \in \mathcal{I}$. Next, introduce the set $\mathcal{O}_M = \mathcal{O}_M(\theta) = \{\boldsymbol{I} \in \mathcal{I} : r^2(\boldsymbol{I}, \theta) \le 2c_3 r^2(\theta) + C_1 M\}$ and the events $\mathcal{A}_M(\boldsymbol{k}) = \big\{\max_{\boldsymbol{J} \in \mathcal{I}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2 \le \frac{5}{\alpha}\rho(\boldsymbol{k}) + C_2 M\big\}$ for $\boldsymbol{k} \in [\boldsymbol{n}]$. The constants $M_1, C_1, C_2 > 0$ are to be chosen later.

Applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{P}_\theta(\|\theta - \hat{\theta}\|^2 \ge \Delta_M) &= \mathbb{P}_\theta\big(\|\theta - \mathrm{P}_{\hat{\boldsymbol{I}}} X\|^2 \ge \Delta_M\big) = \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{I}} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} \ge \Delta_M\Big) \\
&= \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{O}_M} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} \big(1_{\mathcal{A}_M(\boldsymbol{k}(\boldsymbol{I}))} + 1_{\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))}\big) + \sum_{\boldsymbol{I} \in \mathcal{O}_M^c} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} \ge \Delta_M\Big) \\
&\le \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{O}_M} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} 1_{\mathcal{A}_M(\boldsymbol{k}(\boldsymbol{I}))} \ge \tfrac{\Delta_M}{3}\Big) + \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{O}_M} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} 1_{\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))} \ge \tfrac{\Delta_M}{3}\Big) \\
&\quad + \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{O}_M^c} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} \ge \tfrac{\Delta_M}{3}\Big) = T_1 + T_2 + T_3. \qquad (4.23)
\end{aligned}$$

First, let us evaluate T_1. For any $\boldsymbol{I} \in \mathcal{O}_M$, under $\mathcal{A}_M(\boldsymbol{k}(\boldsymbol{I}))$, we have that $\rho(\boldsymbol{k}(\boldsymbol{I})) \le r^2(\boldsymbol{I}, \theta) \le 2c_3 r^2(\theta) + C_1 M$ and

$$\begin{aligned}
R_{\boldsymbol{I}}^2 &= \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2 \le 2c_3 r^2(\theta) + C_1 M + \tfrac{5}{\alpha}\rho(\boldsymbol{k}(\boldsymbol{I})) + C_2 M \\
&\le 2c_3\big(1 + \tfrac{5}{\alpha}\big) r^2(\theta) + \big(C_1 + \tfrac{5}{\alpha} C_1 + C_2\big) M.
\end{aligned}$$

Using this, we derive

$$\begin{aligned}
T_1 &= \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{O}_M} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} 1_{\mathcal{A}_M(\boldsymbol{k}(\boldsymbol{I}))} \ge \tfrac{\Delta_M}{3}\Big) \\
&\le \mathbb{P}_\theta\Big(2c_3\big(1 + \tfrac{5}{\alpha}\big) r^2(\theta) + \big(\tfrac{\alpha+5}{\alpha} C_1 + C_2\big) M \ge \tfrac{\Delta_M}{3}\Big) = 0, \qquad (4.24)
\end{aligned}$$

as $2c_3(1 + \frac{5}{\alpha}) < M_1/3$ and $\frac{\alpha+5}{\alpha} C_1 + C_2 = 1/3$, because we choose $M_1 \triangleq 7c_3(1 + \frac{5}{\alpha})$, $C_1 \triangleq \frac{\alpha}{6(\alpha+5)}$ and $C_2 \triangleq \frac{1}{6}$.

Now we evaluate T_2. Using the Markov and Cauchy-Schwarz inequalities, we obtain

$$\begin{aligned}
T_2 &= \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I} \in \mathcal{O}_M} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} 1_{\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))} \ge \tfrac{\Delta_M}{3}\Big) \\
&\le \frac{\mathbb{E}_\theta \sum_{\boldsymbol{I} \in \mathcal{O}_M} \big(\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 + \|\mathrm{P}_{\boldsymbol{I}}\xi\|^2\big) \hat{p}_{\boldsymbol{I}} 1_{\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))}}{\Delta_M/3} \\
&\le \frac{\sum_{\boldsymbol{I} \in \mathcal{O}_M} \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 \mathbb{P}_\theta\big(\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))\big)}{\Delta_M/3} \\
&\quad + \frac{\sum_{\boldsymbol{I} \in \mathcal{O}_M} \big(\mathbb{E}_\theta \|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big)^{1/2} \big[\mathbb{P}_\theta\big(\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))\big)\big]^{1/2}}{\Delta_M/3}. \qquad (4.25)
\end{aligned}$$

Let us we evaluate the terms in the right hand side of the last inequality. From the definition of the set $\mathcal{O}_M$, it follows that

$$\frac{\|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2}{\Delta_M/3} \le \frac{3(2c_3 r^2(\theta) + C_1 M)}{M_1 r^2(\theta) + M} \le \frac{6c_3}{M_1} + 3C_1 \triangleq C_3 \quad \text{for any } \boldsymbol{I} \in \mathcal{O}_M. \qquad (4.26)$$

Next, $\Delta_M \geq M \geq 1$ as $M \geq 1$. Using this, applying (4.7) with $t_0 = 1/2$ and denoting $C_4 \triangleq \frac{3}{\alpha t_0}$, we obtain that, for any $\boldsymbol{I} \in \mathcal{I}$,

$$\frac{\left[\mathbb{E}_\theta \|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\right]^{1/2}}{\Delta_M/3} \leq \frac{3}{\alpha t_0} \exp\{\rho(\boldsymbol{k}(\boldsymbol{I}))/2\} = C_4 \exp\{\rho(\boldsymbol{k}(\boldsymbol{I}))/2\}. \tag{4.27}$$

By Condition (C1),

$$\begin{aligned}\mathbb{E}_\theta \exp\Big\{\alpha \max_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2\Big\} &\leq \sum_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}}} \mathbb{E}_\theta e^{\alpha\|\mathrm{P}_{\boldsymbol{J}}\xi\|^2} \leq \sum_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}}} e^{\rho(\boldsymbol{k})} \\ &\leq |\mathcal{I}_{\boldsymbol{k}}| e^{\rho(\boldsymbol{k})} \leq e^{2\rho(\boldsymbol{k})},\end{aligned}$$

which, combined the Markov inequality, implies that for any $L \geq 0$

$$\begin{aligned}&\mathbb{P}_\theta\Big(\alpha \max_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2 \geq 2\rho(\boldsymbol{k}) + L\Big) \\ &\quad = \mathbb{P}_\theta\Big(\exp\big\{\alpha \max_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2\big\} \geq \exp\big\{2\rho(\boldsymbol{k}) + L\big\}\Big) \leq e^{-L}.\end{aligned}$$

Applying the last relation for $L = 3\rho(\boldsymbol{k}) + \alpha C_2 M$, we obtain

$$\begin{aligned}\mathbb{P}_\theta(\mathcal{A}_M^c(\boldsymbol{k}(\boldsymbol{I}))) &= \mathbb{P}_\theta\Big(\alpha \max_{\boldsymbol{J}\in\mathcal{I}_{\boldsymbol{k}(\boldsymbol{I})}} \|\mathrm{P}_{\boldsymbol{J}}\xi\|^2 > 5\rho(\boldsymbol{k}(\boldsymbol{I})) + \alpha C_2 M\Big) \\ &\leq \exp\big\{-3\rho(\boldsymbol{k}(\boldsymbol{I})) - \alpha C_2 M\big\}.\end{aligned} \tag{4.28}$$

In view of (4.11), $\sum_{\boldsymbol{I}\in\mathcal{I}} \exp\{-K\rho(\boldsymbol{k}(\boldsymbol{I}))\} \leq (e^K + e^{-K} - 2)^{-1} = \Sigma(K)$ for any $K \geq 1$. Combining this with (4.25), (4.26), (4.27) and (4.28) gives the following bound for T_2:

$$\begin{aligned}T_2 &\leq C_3 \sum_{\boldsymbol{I}\in\mathcal{I}} \exp\{-3\rho(\boldsymbol{k}(\boldsymbol{I})) - \alpha C_2 M\} + C_4 \sum_{\boldsymbol{I}\in\mathcal{I}} \exp\big\{\tfrac{\rho(\boldsymbol{k}(\boldsymbol{I}))}{2} - \tfrac{3\rho(\boldsymbol{k}(\boldsymbol{I}))}{2} - \tfrac{\alpha C_2 M}{2}\big\} \\ &\leq C_3\Sigma(3)e^{-\alpha C_2 M} + C_4\Sigma(1)e^{-\alpha C_2 M/2}.\end{aligned} \tag{4.29}$$

It remains to bound T_3. Applying first the Markov inequality and then the Cauchy-Schwarz inequality, we have

$$\begin{aligned}T_3 = \mathbb{P}_\theta\Big(\sum_{\boldsymbol{I}\in\mathcal{O}_M^c} R_{\boldsymbol{I}}^2 \hat{p}_{\boldsymbol{I}} \geq \tfrac{\Delta_M}{3}\Big) &\leq \frac{\sum_{\boldsymbol{I}\in\mathcal{O}_M^c} \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 \mathbb{E}_\theta \hat{p}_{\boldsymbol{I}}}{\Delta_M/3} \\ &+ \frac{\sum_{\boldsymbol{I}\in\mathcal{O}_M^c} \big(\mathbb{E}_\theta \|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big)^{1/2} \big[\mathbb{E}_\theta \hat{p}_{\boldsymbol{I}}\big]^{1/2}}{\Delta_M/3} = T_{31} + T_{32}.\end{aligned} \tag{4.30}$$

For each $\boldsymbol{I} \in \mathcal{O}_M^c$, we have $c_3 r^2(\theta) \leq r^2(\boldsymbol{I},\theta)/2 - C_1 M/2$, yielding the bound

$$c_2\big(r^2(\boldsymbol{I},\theta) - c_3 r^2(\theta)\big) \geq \frac{c_2}{2} r^2(\boldsymbol{I},\theta) + \frac{c_2 C_1}{2} M.$$

The last relation and Lemma 4.2 entail that, for any $\boldsymbol{I} \in \mathcal{O}_M^c$,

$$\mathbb{E}_\theta \hat{p}_{\boldsymbol{I}} = \mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) \leq \exp\big\{-c_1\rho(\boldsymbol{k}(\boldsymbol{I})) - \tfrac{c_2}{2} r^2(\boldsymbol{I},\theta) - \tfrac{c_2 C_1}{2} M\big\}. \tag{4.31}$$

Now we use the relation (4.31), the fact that $\max_{x\geq 0}\{xe^{-cx}\} \leq (ce)^{-1}$ for any $c>0$, the fact that $\sum_{\boldsymbol{I}\in\mathcal{I}} e^{-c_1\rho(\boldsymbol{k}(\boldsymbol{I}))} \leq \Sigma(c_1)$ (in view of (4.11) as $c_1 \geq 2$) and the trivial relation $\Delta_M \geq M \geq 1$ (as $M \geq 1$), we bound the term T_{31} as follows: with $C_5 \triangleq 6\Sigma(c_1)/(c_2 e)$,

$$\begin{aligned} T_{31} &= \frac{\sum_{\boldsymbol{I}\in\mathcal{O}_M^c} \|\theta - \mathrm{P}_{\boldsymbol{I}}\theta\|^2 \mathbb{E}_\theta \hat{p}_{\boldsymbol{I}}}{\Delta_M/3} \\ &\leq 3 \sum_{\boldsymbol{I}\in\mathcal{O}_M^c} r^2(\boldsymbol{I},\theta) \exp\big\{-c_1\rho(\boldsymbol{k}(\boldsymbol{I})) - \tfrac{c_2}{2} r^2(\boldsymbol{I},\theta) - \tfrac{c_2 C_1}{2} M\big\} \\ &\leq \frac{6e^{-c_2C_1M/2}}{c_2 e} \sum_{\boldsymbol{I}\in\mathcal{O}_M^c} \exp\big\{-c_1\rho(\boldsymbol{k}(\boldsymbol{I}))\big\} \\ &\leq \frac{6\Sigma(c_1)}{c_2 e} e^{-c_2C_1M/2} = C_5 e^{-c_2C_1M/2}. \end{aligned} \tag{4.32}$$

Using the bound (4.7) with $t = c_2/4$, we have $\big[\mathbb{E}_\theta \|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big]^{\frac{1}{2}} \leq \frac{4}{\alpha c_2}\exp\{\frac{c_2}{4}\rho(\boldsymbol{k}(\boldsymbol{I}))\}$ for all $\boldsymbol{I}\in\mathcal{I}$. Besides, $\Delta_M \geq 1$, $r^2(\boldsymbol{I},\theta) \geq \rho(\boldsymbol{k}(\boldsymbol{I}))$ and $\sum_{\boldsymbol{I}\in\mathcal{I}} e^{-c_1\rho(\boldsymbol{k}(\boldsymbol{I}))/2} \leq \Sigma(c_1/2)$ (in view of (4.11) as $c_1 \geq 2$). Piecing all these together with (4.31) and denoting $C_6 \triangleq 12\Sigma(c_1/2)/(\alpha c_2)$, we obtain

$$\begin{aligned} T_{32} &= \frac{\sum_{\boldsymbol{I}\in\mathcal{O}_M^c} \big(\mathbb{E}_\theta \|\mathrm{P}_{\boldsymbol{I}}\xi\|^4\big)^{1/2} \big[\mathbb{E}_\theta \hat{p}_{\boldsymbol{I}}\big]^{1/2}}{\Delta_M/3} \\ &\leq \frac{12e^{-c_2C_1M/4}}{\alpha c_2} \sum_{\boldsymbol{I}\in\mathcal{O}_M^c} \exp\big\{\tfrac{c_2}{4}\rho(\boldsymbol{k}(\boldsymbol{I})) - \tfrac{c_1}{2}\rho(\boldsymbol{k}(\boldsymbol{I})) - \tfrac{c_2}{4} r^2(\boldsymbol{I},\theta)\big\} \\ &\leq \frac{12e^{-c_2C_1M/4}}{\alpha c_2} \sum_{\boldsymbol{I}\in\mathcal{I}} \exp\big\{-\tfrac{c_1}{2}\rho(\boldsymbol{k}(\boldsymbol{I}))\big\} \leq C_6 e^{-c_2C_1M/4}. \end{aligned}$$

Finally, combining (4.23), (4.24), (4.29), (4.30), (4.32) and the last relation, we finish the proof of the theorem: with $M_1 = 7c_3(1+\frac{5}{\alpha})$, $H_1 = C_3\Sigma(3) + C_4\Sigma(1) + C_5 + C_6$ and $m_1 = \min\{\alpha C_2/2, c_2C_1/4\}$,

$$\begin{aligned} \mathbb{P}_\theta\big(\|\hat{\theta}-\theta\|^2 \geq M_1 r^2(\theta) + M\big) &\leq C_3\Sigma(3)e^{-\alpha C_2 M} + C_4\Sigma(1)e^{-\alpha C_2 M/2} \\ &\quad + C_5 e^{-c_2C_1M/2} + C_6 e^{-c_2C_1M/4} \leq H_1 e^{-m_1 M}. \end{aligned}$$

□

Proof of Theorem 4.2. Denote $\mathcal{G}_1 = \mathcal{G}_1(\theta, M) = \{\boldsymbol{I}\in\mathcal{I} : r^2(\boldsymbol{I},\theta) \geq c_3 r^2(\theta) + M\}$, where the constants $c_1 \geq 2$, c_2, c_3 are defined in Lemma 4.2. Applying Lemma 4.2 and (4.11), we obtain

$$\mathbb{P}_\theta\big(\hat{\boldsymbol{I}} \in \mathcal{G}_1\big) = \sum_{\boldsymbol{I}\in\mathcal{G}_1} \mathbb{P}_\theta(\hat{\boldsymbol{I}} = \boldsymbol{I}) \leq e^{-c_2M} \sum_{\boldsymbol{I}\in\mathcal{I}} e^{-c_1\rho(\boldsymbol{k}(\boldsymbol{I}))} \leq \Sigma(c_1) e^{-c_2M},$$

which completes the proof. □

5

ROBUST INFERENCE FOR GENERAL PROJECTION STRUCTURES

Assume that the data (Y, X) come from the model (1.15). The detailed description of the model (1.15) is given in Section 1.5. The goal is to make inference on the parameter θ based on the data (Y, X): recovery of θ (*estimation* and *posterior contraction*), *structure recovery*, and *uncertainty quantification* by constructing an *optimal confidence set*. The detailed description of optimality framework for uncertainty quantification can be found in Section 1.4.

We pursue *local* inference in the sense that no structure on θ is a priori assumed but we aim to extract as much structure (according to the family of structures $\mathcal{I}$, once $\mathcal{I}$ is chosen) as there is in the underlying θ. We will make this notion precise later. We also pursue *robust* inference in the sense that the distribution of ξ is unknown and can depend on θ (often we suppress this dependence in notation), the coordinates ξ_i's of ξ do not have to be iid, even not independent. The distribution of ξ is assumed to satisfy only certain mild condition; see Condition (A1) in Section 5.1. We derive *non-asymptotic* results, which imply asymptotic assertions if needed. Possible asymptotic regimes are: high-dimensional setup $N \to \infty$ (the leading case in the literature for high-dimensional models $\mathcal{Y} = \mathbb{R}^N$), decreasing noise level $\sigma \to 0$, or their combination, e.g., $\sigma = N^{-1/2}$ and $N \to \infty$.

For inference on θ, we exploit the *empirical Bayes approach* and make connection with the *penalization method*. Since any Bayesian approach always delivers also a posterior $\pi(\vartheta|Y)$ (in the posteriors, we will use the variable ϑ to distinguish it from the true parameter θ), an accompanying problem of interest is the contraction of the resulting (empirical Bayes) posterior to the "true" structured θ from the frequentist perspective of the "true" measure $\mathbb{P}_\theta$, which is the distribution of data Y from (1.15). The quality of posterior is characterized by the the posterior contraction rate. We allow this to be a local quantity, i.e., depending on the true θ, while usually in the literature on Bayesian nonparametrics it is a global quantity related to the minimax estimation rates over certain classes. Despite the rapidly growing number of papers about particular high-

dimensional and nonparametric models and structures, there are very few approaches in both frequentist and Bayesian literature that can deal with general classes of high-dimensional and nonparametric models: general posterior contraction rate results are studied in [42, 43, 46, 87], general frameworks for estimation in [42, 55]. We should especially highlight the paper [42] which provided us with important insights for the present study (although our approach is very different). However, all estimation (and posterior contraction) results do not reveal how far an optimal estimator (and posterior) is from the "true" θ. It is of great importance to quantify this uncertainty, which can be cast into the problem of constructing confidence sets for θ.

The scope of this chapter. The keywords summarizing the main novel features of our approach in this chapter are *general*, *robust*, *local*, *refined*, and *EBR-scale*.

The approach is *general* because we develop a general framework of projection structures and study inference problems within this framework by using empirical Bayes and penalization methods. The main inference problem is the *uncertainty quantification*, but on the way we solve the *estimation* problem, *posterior contraction* problem and *(a weak version of) structure recovery* problem as well. As the proposed general framework unifies a broad class of models with various structures, interesting and important on their own right (including graphical/network models), the general framework results deliver a whole avenue of results (many new ones and some known in the literature) for particular models and structures as consequences. There are numerous examples of models and structures falling into our general framework. In Section 5.5 of this chapter, we consider the following models and structures: 1) signal+noise model with smoothness structure (Sobolev ellipsoids and hyperrectangles, analytic and tail classes); 2) signal+noise model under wavelet basis (Besov balls); 3) signal+noise model with (multi-level) sparsity structure (multi-level sparsity is considered for the first time); 4) noisy function on a large graph (Laplacian graph) with smoothness structure; 5) density estimation with smoothness structure; 6) biclustering model (also for stochastic block model and graphon classes); 7) linear regression with sparsity structure, with group sparsity, with group clustering, and with mixture structure; 8) aggregation in nonparametric regression; 9) isotonic, unimodal and convex regressions; 10) dictionary learning; 11) mean matrix with submatrix sparsity; 12) covariance matrix with banding and sparsity structures. Almost all the results on uncertainty quantification problem are new, many known results are improved (obtained local rates improve upon global ones from the literature), some new structures (like multi-level sparsity) are studied for the first time. We emphasize that the scope of our approach extends further than just these specific cases. In fact, the results are readily obtained for any specific model and structure falling into the proposed general framework.

The approach is *robust* in that the model is allowed to be misspecified. Namely, we introduce a family of normal priors, propose an empirical Bayes procedure, and use the normal likelihood, whereas the true model (1.15) does not have to be normal. In fact, the distribution of ξ in (1.15) is not known, the ξ_i's are not necessarily iid (and not even independent), but only satisfying certain mild *exchangeable exponential moment condition*; see Condition (A1) in Section 5.1.

The approach is *local* in that the quality of the inference procedures is measured by

the local quantity, the *oracle rate* $r(\theta)$ (the best rate over a certain family of local rates) which is the best trade-off between the approximation error by a projection structure and the complexity of that approximating projection structure. In a way, it measures the amount of projection structure for each θ: the smaller $r(\theta)$, the more structured θ. To the best of our knowledge, there is no general framework on uncertainty quantification, neither in global nor in local settings for high-dimensional and nonparametric models. We attempt to fill this gap by developing the novel local approach, namely, the radial rate $R(\theta)$ in (1.14) is allowed to be a function of θ, preferably coinciding with the oracle rate $r(\theta)$. The proposed local approach is more powerful than global in that we do not need to impose any specific projection structure, because the local approach automatically exploits the "effective" projection structure of each underlying θ, and our local results imply a whole panorama of global minimax adaptive results over various scales at once, see examples in Section 5.5.

We construct a confidence ball by using the empirical Bayes posterior quantities. Since we want the size of our confidence sets to be of the oracle rate order, this comes with the price that the coverage property can hold uniformly only over some set of parameters satisfying the so called *excessive bias restriction* (EBR) $\Theta_0 = \Theta_{\mathrm{eb}} \subseteq \Theta$. The main result consists in establishing the optimality (1.14) of the constructed confidence ball for the optimality framework $\Theta_0 = \Theta_{\mathrm{eb}}$, $\Theta_1 = \Theta$ and the local radial rate $r(\theta)$. In addition, we also treat the optimality framework with $\Theta_0 = \Theta_1 = \Theta$ in (1.14) by constructing an alternative confidence ball such that its radius is of the order $\sigma N^{1/4} + r(\theta)$. Thus, insisting on the overall uniformity in the coverage and size relations leads to the extra term $\sigma N^{1/4}$ in the expression for the effective radial rate $R(\theta) = r(\theta) + \sigma N^{1/4}$. This fact has also been observed in [67] for the case of linear regression with two sparsity classes. Interestingly, this alternative construction of confidence ball is more preferable for some particular models and structures, e.g., biclustering model (stochastic block model), dictionary learning; see Section 5.5. The point is that, for those models and structures, the extra term $\sigma N^{1/4}$ does not increase the order of the radial rate because $\sigma N^{1/4} \le c r(\theta)$ for the "majority" of θ's, precisely, for all $\theta \in \Theta \backslash \tilde{\Theta}$, with some "thin" set $\tilde{\Theta}$. The set $\tilde{\Theta}$ can be informally described as a set of "highly structured" parameters. This means that, modulo the set $\tilde{\Theta}$ of "highly structured" parameters, there is no deceptiveness issue for those cases. (Speaking informally, these models and structures are already "too difficult" for the term $\sigma N^{1/4}$ to spoil the radial rate.)

The approach is *refined*, namely, we derive the local posterior contraction result for the resulting empirical Bayes posterior $\hat{\pi}(\vartheta|Y)$ in the refined *non-asymptotic exponential probability bound* formulation: $\sup_{\theta\in\Theta} \mathbb{E}_\theta \hat{\pi}(\|\vartheta - \theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2 | Y) \le H_0 e^{-m_0 M}$ for some fixed $M_0, H_0, m_0 > 0$ and arbitrary $M \ge 0$, uniformly in $\theta \in \Theta$. Besides, we obtain the local estimation result for the empirical Bayes posterior mean estimator, also as a non-asymptotic exponential probability bound. This refined formulation provides a rather sharp characterization of the quality of the posterior and the estimator, (finer than, e.g., traditional oracle inequalities in expectation or asymptotic claims for posterior contraction), allowing subtle analysis for various asymptotic regimes. These results, besides being ingredients for the uncertainty quantification problem, are of interest and importance on its own as they establish the local (oracle) optimality of the empirical Bayes posterior and estimator in this refined formulation. As we have mentioned already, the

local results imply in turn the corresponding global (minimax adaptive) results, also in the refined formulation.

Finally, recall that in order to have the targeted optimal radial rate $r(\theta)$ in the size relation of the optimality framework (1.14), the overall uniformity in the coverage property of (1.14) must be sacrificed: $\Theta_0 = \Theta_{\mathrm{eb}} \subseteq \Theta$, where Θ_{eb} is a set of parameters satisfying the so called *excessive bias restriction* (EBR). This set is expected to be of the same type for all models and structures coming from the general framework (1.15): $\Theta_{\mathrm{eb}} = \Theta_{\mathrm{eb}}(t) = \{\theta \in \Theta : b(\theta) \le tV(\theta)\}$, where $b(\theta)$ and $V(\theta)$ are the approximation and complexity (or "bias" and "variance") parts of the squared oracle rate $r^2(\theta) = b(\theta) + V(\theta)$ for the corresponding particular model and structure. It turns out that the EBR leads to a new *EBR-scale* $\{\Theta_{\mathrm{eb}}(t), t \ge 0\}$, which gives a slicing of the entire space: $\Theta = \cup_{t\ge 0}\Theta_{\mathrm{eb}}(t)$. This slicing is very suitable for uncertainty quantification and provides a new perspective at the deceptiveness issue: basically, each parameter θ is deceptive (or non deceptive) to some extent. It is the parameter t that measures the deceptiveness in $\Theta_{\mathrm{eb}}(t)$ and affects the size of the confidence ball needed to provide a guaranteed high coverage uniformly over $\Theta_{\mathrm{eb}}(t)$.

5

This chapter is organized as follows. In Section 5.1 we introduce the notation, the prior, describe the empirical Bayes procedure in detail, make a link with the penalization method, and provide some conditions. Section 5.2, where we also introduce the EBR, contains the main results of the chapter. The proofs of the lemmas and theorems are given in Sections 5.3 and 5.4, respectively. In Section 5.5, we demonstrate how the main general results specify to the above mentioned examples of models and structures in local and minimax settings.

5.1. PRELIMINARIES

First we introduce some notation and notions, then introduce some conditions and a mixture normal prior. Next, by applying the empirical Bayes approach to the normal likelihood (recall that the true model does not have to be normal), we derive an empirical Bayes posterior which we will use in the construction of the estimator and the confidence ball.

At first reading, one may want to skip this section and go ahead to Section 5.2 (one will only need to consult some definitions from Section 5.1) which contains the main results of this chapter.

5.1.1. NOTATION

For a Hilbert space $\mathcal{Y}$, $\langle y, z\rangle$ denotes the scalar product between $y, z \in \mathcal{Y}$. For an $(n_1 \times n_2)$-matrix $x = (x_{ij}) \in \mathbb{R}^{n_1\times n_2}$, we will interchangeably use the same notation x to denote the vector $x = \mathrm{vec}\big[(x_{ij})\big] = (x_{11}, x_{12} \ldots, x_{n_1 n_2})^T$. Conversely, for any $x \in \mathbb{R}^{n_1 n_2}$ we can use matricized indexing $x = (x_{11}, x_{12}, \ldots, x_{n_1 n_2})^T$. Most of the time the vector notation will be used, and it should be clear from the context which notation is meant in each expression.

We will often denote matrices and operators by upright capital letters. The dimensions of matrices and normal distributions should be clear from the context. Throughout we assume the conventions: $|\varnothing| = 0$, $\sum_{I\in\varnothing} a_I = 0$ for any $a_I \in \mathbb{R}$ and $0\log(a/0) = 0$ (hence $(a/0)^0 = 1$) for any $a > 0$.

5.1.2. MULTIVARIATE NORMAL PRIOR

In this section we construct an (empirical) Bayesian procedure and make connection with the penalization method for making inference on θ. Recall that the true parameter θ is assumed to be well approximated by its structured version $\mathrm{P}_{I^*}\theta$ (e.g., it itself can be structured $\theta = \mathrm{P}_{I^*}\theta$), for some "true" structure $I^* \in \mathcal{I}$. The true structure I^* is unknown, so at a later stage we will put a prior on the family of structures $\mathcal{I}$. For now, given a structure I, consider the model $Y = \mathrm{P}_I\theta + \sigma\xi$, approximating the original model (1.15), where P_I is the projection operator onto space $\mathbb{L}_I$, and put first an "unstructured" prior Π on the "unstructured" $\theta \in \Theta$ as follows: $\theta \sim \Pi = \mathrm{N}(\mu, \kappa\sigma^2\mathrm{I})$, where $\kappa = e-1$ and the parameter $\mu \in \mathcal{Y}$ is to be chosen by the empirical Bayes method later. The "unstructured" prior Π on θ leads to the "structured" prior π_I on the "structured" $\theta^I \triangleq \mathrm{P}_I\theta$:

$$\pi_I(\vartheta) = \mathrm{N}(\mathrm{P}_I\mu, \kappa\sigma^2\mathrm{P}_I), \quad I \in \mathcal{I}, \quad \kappa = e-1. \tag{5.1}$$

In this way, we constructed the conditional prior on θ given I: $\theta|I \sim \pi_I(\vartheta)$. The rather specific choice of $\kappa = e-1$ is made for the sake of clean mathematical exposition in later calculations, many other choices are actually possible.

The next very important step in the Bayesian analysis below is that we use the normal likelihood $\ell(\theta, Y) = \bigotimes_i \mathrm{N}(\theta_i, \sigma^2)$, whereas the "true" model $Y \sim \mathbb{P}_\theta$ is not assumed to be normal, but only satisfying Condition (A1). Formally applying Bayesian approach to this prior and the normal likelihood $\ell(\theta, Y)$ delivers the marginal distribution $Y \sim \mathbb{P}_{Y,I} = \mathrm{N}(\mathrm{P}_I\mu, \mathrm{I} + \kappa\mathrm{P}_I)$ and the following posterior distribution on θ:

$$\pi_I(\vartheta|Y) = \mathrm{N}\big(\tfrac{1}{\kappa+1}\mathrm{P}_I\mu + \tfrac{\kappa}{\kappa+1}\mathrm{P}_I Y, \tfrac{\kappa\sigma^2}{\kappa+1}\mathrm{P}_I\big). \tag{5.2}$$

Note that in general the covariance matrix in (5.1) is not invertible, but the Bayes formula for the conjugate normal-normal model still holds with the Moore-Penrose inverse P_I^- of P_I instead of the usual inverse (recall that $\mathrm{P}^- = \mathrm{P}$ for any projection operator P).

Let us now put a prior on I:

$$\lambda_I = c_\varkappa e^{-\varkappa\rho(s(I))}, \quad I \in \mathcal{I}, \tag{5.3}$$

where $c_\varkappa$ is the normalizing constant (i.e., $\sum_{I\in\mathcal{I}}\lambda_I = 1$), $\rho(s)$ satisfies (1.21), the mapping $s(I)$ is from Condition (A1), the parameter $\varkappa$ satisfies the relation

$$\varkappa > \bar{\varkappa} \triangleq (32\nu + 10 + \alpha)/(4\alpha), \tag{5.4}$$

α and ν are from Conditions (A1) and (A2), respectively.

Combining (5.1) and (5.3) gives the mixture prior on θ: $\pi = \sum_{I\in\mathcal{I}}\lambda_I\pi_I$. This leads to the marginal distribution of Y: $\mathbb{P}_Y = \sum_{I\in\mathcal{I}}\lambda_I\mathbb{P}_{Y,I}$, $\mathbb{P}_{Y,I} = \mathrm{N}(\mathrm{P}_I\mu, \sigma^2(\mathrm{I}+\kappa\mathrm{P}_I))$, where the density of the distribution $\mathbb{P}_{Y,I} = \mathrm{N}(\mathrm{P}_I\mu, \sigma^2(\mathrm{I}+\kappa\mathrm{P}_I))$ is

$$\varphi(y, \mathrm{P}_I\mu, \sigma^2(\mathrm{I}+\kappa\mathrm{P}_I)) = \frac{e^{-(y-\mathrm{P}_I\mu)^T(\mathrm{I}-\frac{\kappa}{\kappa+1}\mathrm{P}_I)(y-\mathrm{P}_I\mu)/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}(1+\kappa)^{\dim(\mathbb{L}_I)/2}}, \tag{5.5}$$

because $(\mathrm{I}+\kappa\mathrm{P})^{-1} = \mathrm{I} - \frac{\kappa}{\kappa+1}\mathrm{P}$, $\det(\mathrm{I}+\kappa\mathrm{P}) = (1+\kappa)^{\mathrm{rank}(\mathrm{P})}$ for any projection operator P, and $\mathrm{rank}(\mathrm{P}_I) = \dim(\mathbb{L}_I)$. The posterior of θ becomes

$$\pi(\vartheta|Y) = \pi_\varkappa(\vartheta|Y) = \sum_{I\in\mathcal{I}}\pi(\vartheta, I|Y) = \sum_{I\in\mathcal{I}}\pi(\vartheta|Y, I)\pi(I|Y), \tag{5.6}$$

where $\pi(\vartheta|Y,I)=\pi_I(\vartheta|Y)$ is defined by (5.2) and the posterior for I is

$$\pi(I|Y)=\frac{\lambda_I \mathbb{P}_{Y,I}}{\sum_{J\in\mathcal{I}}\lambda_J \mathbb{P}_{Y,J}}. \tag{5.7}$$

5.1.3. EMPIRICAL BAYES POSTERIOR

The parameter μ is yet to be chosen in the prior. We apply the empirical Bayes approach. The marginal likelihood P_Y is readily maximized with respect to μ: $\operatorname{argmin}_\mu\{(Y-\mathrm{P}_I\mu)^T(\mathrm{I}-\frac{\kappa}{\kappa+1}\mathrm{P}_I)(Y-\mathrm{P}_I\mu)\}=Y$. Substituting Y instead of μ in the expressions (5.2), (5.6) and (5.7) yields the empirical Bayes posterior

$$\tilde{\pi}(\vartheta|Y)=\tilde{\pi}_\varkappa(\vartheta|Y)=\sum_{I\in\mathcal{I}}\tilde{\pi}(\vartheta|Y,I)\tilde{\pi}(I|Y), \tag{5.8}$$

called *empirical Bayes model averaging* (EBMA) posterior, where the EBMA posterior for θ given I is

$$\tilde{\pi}(\vartheta|Y,I)=\tilde{\pi}_I(\vartheta|Y)=\mathrm{N}\big(\mathrm{P}_I Y,\tfrac{\kappa\sigma^2}{\kappa+1}\mathrm{P}_I\big) \tag{5.9}$$

and the empirical Bayes posterior for I is

$$\tilde{\pi}(I|Y)=\tilde{\pi}_I=\frac{\lambda_I\exp\{-\frac{1}{2\sigma^2}[\|(\mathrm{I}-\mathrm{P}_I)Y\|^2+\sigma^2\dim(\mathbb{L}_I)]\}}{\sum_{J\in\mathcal{I}}\lambda_J\exp\{-\frac{1}{2\sigma^2}[\|(\mathrm{I}-\mathrm{P}_J)Y\|^2+\sigma^2\dim(\mathbb{L}_J)]\}}. \tag{5.10}$$

When deriving (5.10), we used (5.5), $\kappa=e-1$ and the fact that $(\mathrm{I}-\mathrm{P})(\mathrm{I}-\frac{\kappa}{1+\kappa}\mathrm{P})(\mathrm{I}-\mathrm{P})=(\mathrm{I}-\mathrm{P})$ for any projection operator P. Let $\tilde{\mathbb{E}}$ and $\tilde{\mathbb{E}}_I$ be the expectations with respect to the EBMA measures $\tilde{\pi}(\vartheta|Y)$ and $\tilde{\pi}_I(\vartheta|Y)$, respectively. Then $\tilde{\mathbb{E}}_I(\vartheta|Y)=\mathrm{P}_I Y$, $I\in\mathcal{I}$. Introduce the *EBMA posterior mean* estimator

$$\tilde{\theta}=\tilde{\mathbb{E}}(\vartheta|Y)=\sum_{I\in\mathcal{I}}\tilde{\mathbb{E}}_I(\vartheta|Y)\tilde{\pi}(I|Y)=\sum_{I\in\mathcal{I}}(\mathrm{P}_I Y)\tilde{\pi}(I|Y). \tag{5.11}$$

Consider yet alternative empirical Bayes posterior. First derive an empirical Bayes structure selector $\hat{I}$ by maximizing $\tilde{\pi}(I|Y)$ over $I\in\mathcal{I}$. This boils down to

$$\hat{I}=\operatorname*{argmax}_{I\in\mathcal{I}}\tilde{\pi}(I|Y)=\operatorname*{argmin}_{I\in\mathcal{I}}\big\{\|Y-\mathrm{P}_I Y\|^2+\sigma^2\mathrm{pen}(I)\big\}, \tag{5.12}$$

which is essentially the *penalization method* with the penalty $\mathrm{pen}(I)=2\varkappa\rho(s(I))+\dim(\mathbb{L}_I)$. Plugging in this into $\tilde{\pi}_I(\vartheta|Y)$ defined by (5.9) gives the corresponding empirical Bayes posterior (called *empirical Bayes model selection* (EBMS) posterior), yielding also the *EBMS mean estimator* for θ:

$$\check{\pi}(\vartheta|Y)=\tilde{\pi}_{\hat{I}}(\vartheta|Y)=\mathrm{N}\big(\mathrm{P}_{\hat{I}}Y,\tfrac{\kappa\sigma^2}{\kappa+1}\mathrm{P}_{\hat{I}}\big),\quad \check{\theta}=\check{\mathbb{E}}(\vartheta|Y)=\mathrm{P}_{\hat{I}}Y, \tag{5.13}$$

where $\check{\mathbb{E}}$ denotes the expectation with respect to the EBMS measure $\check{\pi}(\vartheta|Y)$. Notice that, like (5.8), $\check{\pi}(\vartheta|Y)$ defined by (5.13) can also be seen formally as mixture

$$\check{\pi}(\vartheta|Y)=\tilde{\pi}_{\hat{I}}(\vartheta|Y)=\sum_{I\in\mathcal{I}}\tilde{\pi}_I(\vartheta|Y)\check{\pi}(I|Y),\quad \check{\pi}(I|Y)=1\{I=\hat{I}\}, \tag{5.14}$$

where the mixing distribution $\check{\pi}(I|Y)=1\{I=\hat{I}\}$, the empirical Bayes posterior for I, is degenerate at $\hat{I}$.

Remark 5.1. *In a way, the EBMA posterior $\tilde{\pi}(\vartheta|Y)$ defined by (5.8) is "more Bayesian" than the EBMS posterior $\check{\pi}(\vartheta|Y)$ defined by (5.13). Note however that, while the penalization method gives only an estimator, the EBMS method does also yield a posterior.*

From now on, by $\hat{\pi}(\vartheta|Y)$ we denote either $\tilde{\pi}(\vartheta|Y)$ defined by (5.8) or $\check{\pi}(\vartheta|Y)$ defined by (5.13), and $\hat{\theta}$ will stand either for $\tilde{\theta}$ defined by (5.11) or for $\check{\theta}$ defined by (5.13). In case $\hat{\pi}(I|Y) = \check{\pi}(I|Y) = 1\{I = \hat{I}\}$, the meaning of $\hat{\pi}(I \in \mathcal{G}|Y)$ for any $\mathcal{G} \subseteq \mathcal{I}$ is as follows: $\hat{\pi}(I \in \mathcal{G}|Y) = \check{\pi}(I \in \mathcal{G}|Y) = 1\{\hat{I} \in \mathcal{G}\}$ so that $\mathbb{E}_\theta \check{\pi}(I \in \mathcal{G}|Y) = \mathbb{P}_\theta(\hat{I} \in \mathcal{G})$.

5.2. MAIN RESULTS

In this section we present the main results of this chapter. Throughout we assume that Conditions (A1), (A2) and (A3) (defined in Chapter 1) are fulfilled.

5.2.1. ORACLE RATE

For $I \in \mathcal{I}$, consider the projection estimator $\mathrm{P}_I Y$ for estimating θ. By Condition (A1) and Jensen's inequality, we obtain the following upper bound for the estimator $\mathrm{P}_I Y$: for some $C > 0$,

$$\mathbb{E}_\theta \|\theta - \mathrm{P}_I Y\|^2 = \|\theta - \mathrm{P}_I \theta\|^2 + \sigma^2 \mathbb{E}_\theta \|\mathrm{P}_I \xi\|^2 \le \|\theta - \mathrm{P}_I \theta\|^2 + C\sigma^2 d_{s(I)}.$$

Ideally, we would like to mimic the local rate for the best (oracle) choice of the projection structure $\min_{I \in \mathcal{I}} \left(\|\theta - \mathrm{P}_I \theta\|^2 + \sigma^2 d_{s(I)}\right)$, uniformly in $\theta \in \Theta$. However, as is shown for some particular models, this is impossible unless we use a majorant $\rho(s(I)) \ge d_{s(I)} + \log|\mathcal{I}_{s(I)}|$ instead of just $d_{s(I)}$. The layer complexity term $\log|\mathcal{I}_{s(I)}|$ is the "price" for not knowing the structure. This motivates the following definition. Introduce the family of local rates

$$r^2(I,\theta) = \|\theta - \mathrm{P}_I \theta\|^2 + \sigma^2 \rho(s(I)), \quad I \in \mathcal{I},$$

for some $\rho(s)$ satisfying (1.21) and the slicing mapping $s(I)$ from Condition (A1). For each θ there exists the best structure $I_o = I_o(\theta) = I_o(\theta, \sigma^2)$ (if not unique, take any minimizer) corresponding to the fastest local rate

$$r^2(\theta) = \min_{I \in \mathcal{I}} r^2(I,\theta) = r^2(I_o,\theta) = \|\theta - \mathrm{P}_{I_o} \theta\|^2 + \sigma^2 \rho(s(I_o)), \tag{5.15}$$

representing the optimal trade-off between the approximation term $\|\theta - \mathrm{P}_{I_o}\theta\|^2$ and the complexity term $\rho(s(I_o))$ satisfying (1.21). We call I_o by *oracle structure* (or just *oracle*) and the quantity $r^2(\theta)$ by *oracle rate*.

Remark 5.2. *Often we will have that $d_{s(I)} = \dim(\mathbb{L}_I) = d_I$ and $\rho(s(I)) = d_I + \log|\mathcal{I}_{s(I)}|$. This is the case in many particular models and structures that we consider in Section 5.5. If $I^* \in \mathcal{I}$ is the true structure, i.e., $\theta \in \mathbb{L}_{I^*}$, $s^* = s(I^*)$ and $\mathcal{I}_{s^*} = \{I^*\}$, then, by the oracle definition (5.15) and the facts that $\|\theta - \mathrm{P}_{I^*}\theta\|^2 = 0$ and $\mathbb{L}_{I^*} \subseteq \mathbb{R}^N$, we have*

$$r^2(\theta) \le r^2(I^*,\theta) = \sigma^2 \rho(s(I^*)) = \sigma^2 d_{s(I^*)} = \sigma^2 \dim(\mathbb{L}_{I^*}) \le N\sigma^2. \tag{5.16}$$

If such a true structure does not exist, we can assume without loss of generality that there is an $\bar{I} \in \mathcal{I}$ such that $\mathbb{L}_{\bar{I}} = \mathbb{R}^N$, $\bar{s} = s(\bar{I})$ and $\mathcal{I}_{\bar{s}} = \{\bar{I}\}$. This would lead again to the bound (5.16): $r^2(\theta) \le r^2(\bar{I},\theta) = \sigma^2 \dim(\mathbb{L}_{\bar{I}}) = N\sigma^2$. This is of course not surprising as the oracle performance should not be worse than the simplistic procedure $\hat{\theta} = Y$.

Remark 5.3. *Suppose we have two different family of structures $\mathcal{I}_1$ and $\mathcal{I}_2$, with corresponding (different) families of linear spaces $\mathbb{L}_I$ and (different) majorants. We say that the family $\mathcal{I}_1$* covers *the family $\mathcal{I}_2$ if for any $I \in \mathcal{I}_2$ there exists $I' = I'(I) \in \mathcal{I}_1$ such that $r^2(I',\theta) \le r^2(I,\theta)$ for all $\theta \in \Theta$ (up-to-a-constant relation will do as well). If $\mathcal{I}_1$ covers $\mathcal{I}_2$, there is no point in considering the family $\mathcal{I}_2$, one should use the family $\mathcal{I}_1$. The family $\mathcal{I}_1$ and the family $\mathcal{I}_2$ covered by $\mathcal{I}_1$ could be of very different natures. But sometimes $\mathcal{I}_1$ can be a subfamily of $\mathcal{I}_2$. This happens when a chunk of structures in $\mathcal{I}_2$ can be dominated by just one structure. Then we can remove those structures without any harm, obtaining a new adjusted family $\mathcal{I}_1$. The complexity term in the majorant gets adjusted for some $I \in \mathcal{I}_1$, leading to an* elbow effect *in the rate and improving the resulting oracle rate. We will see how this* elbow effect *is exhibited in several examples of Section 5.5.*

5.2.2. ESTIMATION AND CONTRACTION RESULTS WITH ORACLE RATE

Recall the quantities: the empirical Bayes posterior $\hat\pi(\vartheta|Y)$, which is either EBMA posterior $\tilde\pi(\vartheta|Y)$ defined by (5.8) or EBMS posterior $\check\pi(\vartheta|Y)$ defined by (5.13); the empirical Bayes posterior mean $\hat\theta$, which is either $\tilde\theta$ defined by (5.11) or $\check\theta$ defined by (5.13); and the oracle rate $r(\theta)$ defined by (5.15). The following theorem establishes that the empirical Bayes posterior $\hat\pi(\vartheta|Y)$ contracts (from the frequentist $\mathbb{P}_\theta$-perspective) to θ with the oracle rate $r(\theta)$, and the empirical Bayes posterior mean $\hat\theta$ converges to θ with the oracle rate $r(\theta)$, uniformly over the entire parameter space.

Theorem 5.1. *Let Conditions (A1) and (A2) be fulfilled. Then there exist constants $M_0, M_1, H_0, H_1, m_0, m_1 > 0$ such that for any $\theta \in \Theta$ and any $M \ge 0$,*

$$\mathbb{E}_\theta \hat\pi\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2 | Y\big) \le H_0 e^{-m_0 M}, \tag{5.17}$$

$$\mathbb{P}_\theta\big(\|\hat\theta-\theta\|^2 \ge M_1 r^2(\theta) + M\sigma^2\big) \le H_1 e^{-m_1 M}. \tag{5.18}$$

The constants in the theorem depend only on α and some also on $\varkappa$, the exact expressions can be found in the proof.

Remark 5.4. *Notice that already claim (5.17) of Theorem 5.1 contains an oracle bound for the estimator $\hat\theta$. Indeed, by Jensen's inequality, we get the oracle inequality in expectation:*

$$\mathbb{E}_\theta\|\hat\theta-\theta\|^2 \le \mathbb{E}_\theta\hat{\mathbb{E}}(\|\vartheta-\theta\|^2|Y) \le M_0 r^2(\theta) + H_0\int_0^{+\infty} e^{-m_0 u/\sigma^2}\,du = M_0 r^2(\theta) + \tfrac{H_0\sigma^2}{m_0}. \tag{5.19}$$

Similarly we can show that also (5.18) implies (5.19). This means that claim (5.18) is actually stronger than (5.19) and therefore requires a separate proof.

Remark 5.5. *The non-asymptotic exponential probability bounds in the both claims of the theorem provide a very refined characterization of the quality of the posterior $\hat\pi(\vartheta|Y)$ and estimator $\hat\theta$, finer than, e.g., the traditional oracle inequalities in expectation like (5.19) (since (5.19) follows from (5.18), see Remark 5.4). This refined formulation allows for subtle analysis in various asymptotic regimes ($N\to\infty$, $\sigma\to 0$, or their combination) as we can let M depend in any way on N, σ, or both.*

Now we give several technical definitions which we will need in the claims. For the constants α from Condition (A1) and $\varkappa$ from (5.3), define

$$\bar\tau = \bar\tau(\varkappa,\alpha) \triangleq 3(1+\varkappa\alpha+\alpha/2)/\alpha. \tag{5.20}$$

Next, for some $\delta \in (0,1)$, fix some $\tau_0 = \tau_0(\delta)$ such that $\tau_0 > \frac{1+\delta}{1-\delta}\bar{\tau}$, where $\bar{\tau}$ is defined by (5.20). For example, take $\delta = 0.1$ and $\tau_0 = \frac{11}{9}\bar{\tau} + 0.1$. For this τ_0 and any $\theta \in \Theta$, define

$$I_* = I_*(\theta) = I_*(\theta, \delta) \triangleq I_o(\theta, \tau_0 \sigma^2) = I_o^{\tau_0}, \tag{5.21}$$

where $I_o(\theta, \sigma^2)$ is defined by (5.15). We call the quantity $I_o^\tau = I_o(\theta, \tau\sigma^2)$, for $\tau \geq 0$, by *τ-oracle*, which is just the oracle defined by (5.15) with σ^2 substituted by $\tau\sigma^2$. Notice that $\rho(s(I_o^{\tau_1})) \geq \rho(s(I_o^{\tau_2}))$ for $\tau_1 \leq \tau_2$. All τ-oracle rates are related to the oracle rate by the trivial relations: $r^2(\theta) \leq r^2(I_o^\tau, \theta) \leq \tau r^2(\theta)$ for $\tau \geq 1$, and $r^2(I_o^\tau, \theta) \leq r^2(\theta) \leq \tau^{-1} r^2(I_o^\tau, \theta)$ for $0 < \tau < 1$.

When proving the above theorem, as byproduct we also obtain a result about the frequentist behavior of the structure selector $\hat{I}$ and the empirical Bayes posterior for I, saying basically that $\hat{I}$ and $\hat{\pi}(I|Y)$ "live" on a set that is, in a sense, almost as good as the oracle I_o.

Theorem 5.2. *Let Conditions* (A1) *and* (A2) *be fulfilled, ν, C_ν be from Condition* (A2)*. The following relations hold for any $\theta \in \Theta$ and $M \geq 0$.*

(i) Let c_1, c_2, c_3 be the constants defined in Lemma 5.2. Then

$$\mathbb{E}_\theta \hat{\pi}(I \in \mathcal{I} : r^2(I, \theta) \geq c_3 r^2(\theta) + M\sigma^2 \big| Y) \leq C_\nu e^{-c_2 M}.$$

(ii) Let $\varkappa \geq \alpha^{-1}\nu$ (implied by (5.4)) and Condition (A3) *be fulfilled. Then there exists $m_1' > 0$ such that*

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : \rho(s(I)) \leq \delta \rho(s(I_*)) - M \big| Y\big) \leq C_\nu e^{-m_1' M}, \tag{5.22}$$

where $I_* = I_*(\theta, \delta)$ is defined by (5.21).

(iii) Let $\varkappa \geq \frac{2\nu + 2\alpha + 3}{2\alpha}$ (implied by (5.4)) and Condition (A3') *be fulfilled (given in Remark 1.9). Then there exists $M_0' > 0$ such that*

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : \rho(s(I)) \geq M_0' \rho(s(I_o)) + M \big| Y\big) \leq C_\nu e^{-M/2}.$$

We can interpret the above theorem as *structure recovery*, but in a somewhat weak sense. Namely, Theorem 5.2 says basically that the empirical Bayes posterior $\hat{\pi}(I|Y)$ and the structure selector $\hat{I}$ "live" in the set of structures resembling the oracle structure I_o, in the sense that the rates and complexities for the structures from this set are in a proximity of the oracle rate and complexity, respectively. Recall that in general the oracle structure is not the same as the so called true structure.

5.2.3. CONFIDENCE BALL UNDER EXCESSIVE BIAS RESTRICTION (EBR)

Theorem 5.1 establishes strong local optimal properties of the empirical Bayes posterior $\hat{\pi}(\vartheta|Y)$ and mean $\hat{\theta}$, but this is not enough to solve the uncertainty quantification problem yet. As a first candidate for confidence ball, let us construct a credible ball by using the EBMS posterior $\hat{\pi}(\vartheta|Y) = \check{\pi}(\vartheta|Y)$ defined by (5.13). As $\check{\pi}(\vartheta|Y) = \mathrm{N}\big(\check{\theta}, \frac{\kappa}{\kappa+1}\sigma^2 \mathrm{P}_{\hat{J}}\big)$, denoting by $\chi^2_{k,\alpha}$ the $(1-\alpha)$-quantile of χ^2_k-distribution, we have

$$\check{\pi}\big(\|\vartheta - \check{\theta}\|^2 \leq \sigma^2 \chi^2_{\dim(\mathbb{L}_{\hat{J}}),\alpha} \big| Y\big) \geq \check{\pi}\big(\|\vartheta - \check{\theta}\|^2 \leq \tfrac{\kappa}{\kappa+1}\sigma^2 \chi^2_{\dim(\mathbb{L}_{\hat{J}}),\alpha} \big| Y\big) = 1 - \alpha.$$

Since $\chi^2_{\dim(\mathbb{L}_{\hat{I}}),\alpha}$ is bounded by a constant multiple of $\dim(\mathbb{L}_{\hat{I}})$, for simplicity the latter can replace the former to obtain a credible ball. Then $B(\check{\theta}, M\sigma[\dim(\mathbb{L}_{\hat{I}})]^{1/2})$ is a credible ball for θ, which can be guaranteed to have a given level of credibility by choosing a sufficiently large M. However, $B(\check{\theta}, M\sigma[\dim(\mathbb{L}_{\hat{I}})]^{1/2})$ cannot have a guaranteed coverage, since otherwise in some particular models (cf. [13]) this would mean that the estimator $\check{\theta}$ would converge to θ uniformly in $\theta \in \Theta$ at the smaller oracle rate with $\tilde{\rho}(s(I_o)) = \dim(\mathbb{L}_{I_o})$ instead of $\rho(s(I_o)) = d_{s(I_o)} + \log|\mathcal{I}_{s(I_o)}|$. But $\log|\mathcal{I}_{s(I_o)}|$ can be the dominating term in $\rho(s(I_o))$, e.g., for sparsity structures (see [13] and Section 5.5). This would contradict the lower bounds from the literature. Basically, the posterior $\check{\pi}(\vartheta|Y)$ is well concentrated (in fact, " too concentrated"), but not around the truth, rather around its mean $\check{\theta}$ which in general is away from the truth by the distance of a bigger order than the concentration rate. Hence, to obtain coverage, the radius of confidence balls must be at least of the order $\rho(s(I_o)) = d_{s(I_o)} + \log|\mathcal{I}_{s(I_o)}|$. Of course, the oracle structure I_o is not known, but we have the structure selector $\hat{I}$ defined by (5.12), which is in a way close to the oracle structure, according to Theorem 5.2.

The above heuristics suggests to use $\rho(s(\hat{I}))$ as a proxy for $\rho(s(I_o))$. According to Theorem 5.2, $\hat{I}$ lives in the "complexity shell" $\delta\sigma^2\rho(s(I_*)) - M\sigma^2 \le \sigma^2\rho(s(\hat{I})) \le c_3 r^2(\theta) + M\sigma^2$ with a large probability. So, if we want the size of confidence ball to be not of a bigger order than oracle rate, it seems reasonable to use the following data dependent radius

$$\hat{r}^2 = \hat{r}^2(Y) = \sigma^2 + \sigma^2\rho(s(\hat{I})). \tag{5.23}$$

We will show that the size property holds for the radial rate equal to the oracle rate, uniformly over $\theta \in \Theta$. But then there is an inevitable problem with coverage: the coverage property does not hold uniformly. Indeed, the complexity shell can be too wide if $\sigma^2\rho(s(I_*)) \ll r^2(\theta)$. If this happens (for deceptive θ's), then the coverage property of a ball with radius of order $\hat{r}$ cannot be guaranteed because its radius can be of a smaller order than the oracle rate $r^2(\theta)$. This problem will not occur for those θ's (called non-deceptive) for which the bias part of the oracle rate is within a multiple of its variance part. This discussion motivates introducing the following condition.

CONDITION EBR. We say that $\theta \in \Theta$ satisfies the *excessive bias restriction* (EBR) condition with structural parameter $t \ge 0$ if $\theta \in \Theta_{\mathrm{eb}}(t)$, where the corresponding set (called the *EBR class*) is

$$\Theta_{\mathrm{eb}}(t) = \Theta_{\mathrm{eb}}(t, \tau_0) = \bigl\{\theta \in \Theta : \|\theta - \mathrm{P}_{I_*}\theta\|^2 \le t\sigma^2\bigl(1 + \rho(s(I_*))\bigr)\bigr\}, \tag{5.24}$$

where the τ_0-oracle structure $I_* = I_o(\theta, \tau_0\sigma^2)$ is defined by (5.21). The condition EBR essentially requires that the bias part of the τ_0-oracle rate $r^2(I_*, \theta)$ is dominated by a multiple of its variance part (additional σ^2 is needed to handle the case $\rho(s(I_*)) = 0$). Clearly, $\Theta_{\mathrm{eb}}(t_1) \subseteq \Theta_{\mathrm{eb}}(t_2)$ for $t_1 \le t_2$.

Now we use the center $\hat{\theta}$ and the radius $\hat{r}$ to construct a confidence ball for θ. The following theorem describes the coverage and size properties of the confidence ball based on $\hat{\theta}$ and $\hat{r}$.

Theorem 5.3. *Let Conditions* (A1), (A2) *and* (A3) *be fulfilled,* $\Theta_{\mathrm{eb}}(t)$ *be defined in* (5.24). *Then there exist constants* $M_2, M_3, H_2, H_3, m_2, m_3 > 0$ *such that for any* $t, M \ge 0$, *and with*

$$\hat{R}_M^2 = \hat{R}_M^2(M_2) = (t+1)M_2\hat{r}^2 + (t+2)M\sigma^2,$$

$$\sup_{\theta\in\Theta_{\rm eb}(t)} \mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, \hat{R}_M)\big) \le H_2 e^{-m_2 M}, \qquad \mathbb{P}_\theta\big(\hat{r}^2 \ge M_3 r^2(\theta) + (M+1)\sigma^2\big) \le H_3 e^{-m_3 M}.$$

The size relation (second relation) holds without Condition (A3) and is uniform in $\theta \in \Theta$.

Moreover, if, instead of Condition (A3)*, stronger Condition* (A3')*is fulfilled, then a stronger version of the size relation holds:* $\mathbb{P}_\theta\big(\hat{r}^2 \ge M_0'\rho(s(I_o))\sigma^2 + (M+1)\sigma^2\big) \le C_v e^{-M/2}$, *where the constants* M_0' *and* C_v *are from Theorem 5.2.*

Remark 5.6. *Recall that* $I_* = I_o(\theta, \tau_0\sigma^2)$ *from* (5.24) *is the* τ_0*-oracle. It may be desirable to impose an EBR condition in terms of the "standard" oracle* $I_o = I_o(\theta, \sigma^2)$ *rather than the* τ_0*-oracle. By rewriting the original model* (1.15) *as* $Y\tau_0^{-1/2} = \theta\tau_0^{-1/2} + \sigma\tau_0^{-1/2}\xi$, *it is not difficult to see that we can construct a confidence ball with the radius* $\tau_0\hat{R}_M$ *satisfying the coverage property as above, but now uniformly over* $\Theta_{\rm eb}(t, 1)$.

Remark 5.7. *When proving the coverage relation of Theorem 5.3, we actually established the following uniform local assertion: there exist constants* $M_2, \alpha_1, m_1'', H_2, m_2 > 0$ *such that for any* $\theta \in \Theta$ *and any* $M \ge 0$,

$$\begin{aligned}\mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, [(b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2]^{1/2}\big)\\ \le H_1 e^{-m_1 M} + C_v e^{-m_1'' M} \le H_2 e^{-m_2 M},\end{aligned} \tag{5.25}$$

where the constants H_1, m_1 *are defined in Theorem 5.1,* C_v *is from Condition* (A2)*, and the quantity* $b(\theta)$ *(called* excessive bias ratio*) is defined by*

$$b(\theta) = b(\theta, \tau_0) = \frac{\|\theta - \mathrm{P}_{I_*}\theta\|^2}{\sigma^2 + \sigma^2\rho(s(I_*))}. \tag{5.26}$$

Although the newly formulated coverage relation (5.25) *is now uniform over the entire space* $\theta \in \Theta$, *the main (and unavoidable) problem is its dependence on* $b(\theta)$. *That is why we introduced the EBR condition which essentially provides control over the quantity* $b(\theta)$*: indeed,* $\Theta_{\rm eb}(t) = \{\theta \in \Theta : b(\theta) \le t\}$.

Remark 5.8. *The smaller constant* τ_0 *(involved in the definition of the EBR condition) is, the less restrictive the EBR condition is, the limiting case* $\tau_0 \downarrow 0$ *corresponds basically to no condition. We treat a general situation, with only Condition* (A1) *assumed for* ξ, *so that we have a lower bound for* τ_0 *in terms of* α *which is possibly too conservative for each specific distribution of* ξ. *However, even for any specific distribution of* ξ, *the value of the constant* $\tau_0 > 0$ *in the EBR condition is always bounded away from zero (further from zero for "bad"* ξ*'s).*

Remark 5.9. *The EBR leads to the new* EBR-scale $\{\Theta_{\rm eb}(t), t \ge 0\}$ *which gives a slicing of the entire space* $\Theta = \cup_{t\ge 0}\Theta_{\rm eb}(t)$. *This slicing is very suitable for uncertainty quantification and provides a new perspective at the deceptiveness issue (discussed in Chapter 1): basically, each parameter* θ *is deceptive (or non deceptive) to some extent. It is the parameter* t *that measures the deceptiveness in* $\Theta_{\rm eb}(t)$ *and affects the size of the confidence ball needed to provide a guaranteed high coverage uniformly over* $\Theta_{\rm eb}(t)$.

5.2.4. CONFIDENCE BALL OF $N^{1/4}$-RADIUS WITHOUT EBR

Suppose we want to construct a confidence ball of a full coverage uniformly over the whole space Θ. Recall however that for "signal+noise" models, in view of the negative results of [7, 28, 60, 67] mentioned in Chapter 1, no data dependent ball can have uniform coverage and adaptive size simultaneously. When insisting on the uniform coverage, one must have an additional term of the order $\sigma N^{1/4}$ in the radial radius. Let us give a heuristics behind this. An idea is to mimic the quantity $\|\theta-\hat\theta\|^2$ by $\hat R^2=\|Y-\hat\theta\|^2$. Clearly, there is a lot of bias in $\hat R^2$, the biggest part of which is due to the term $\sigma^2\|\xi\|^2$ contained in $\hat R$. To de-bias for that part, we need to subtract its expectation $\sigma^2\mathbb{E}\|\xi\|^2$. However, even the de-biased version of $\hat R^2$ can only be controlled up to a margin of the order $\sigma^2\sqrt{N}$. That is why a term of the order $\sigma N^{1/4}$ is necessary in the radius of the confidence ball to provide coverage uniformly over the whole space Θ.

To handle some technical issues, we impose the following condition.

CONDITION (A4). Besides Y given by (1.15), we also observe $Y'=\theta+\sigma\xi'$ independent of Y, where the random vector ξ' satisfies the following relations:

$$\begin{aligned}&\mathbb{P}\big(|\langle v,\xi'\rangle|\ge\sqrt{M}\big)\le\psi_1(M)\ \ \forall\, v\in\mathbb{R}^N:\|v\|=1;\\&\mathbb{P}\big(\big|\|\xi'\|^2-V(Y')\big|\ge M\sqrt{N}\big)\le\psi_2(M),\quad \text{for some statistic } V(Y').\end{aligned}\tag{A4}$$

Here $\psi_1(M),\psi_2(M)$ are some positive monotonically decreasing functions such that $\psi_1(M)\downarrow 0$ and $\psi_2(M)\downarrow 0$ as $M\uparrow\infty$.

Remark 5.10. *Typically, $\mathbb{E}\xi'_i=0$, $\mathrm{Var}(\xi'_i)=1$, $i\in[N]$, then $V(Y')=N$. Condition (A4) is satisfied for independent normals $\xi_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)$ even if we do not have the sample Y' at our disposal. Indeed, in this case we can "duplicate" the observations by randomization at the cost of doubling the variance in the following manner: create samples $Y'=Y+\sigma Z$ and $Y''=Y-\sigma Z$, for a $Z=(Z_1,\dots,Z_N)$ (independent of Y) such that $Z_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)$. Relations (A4) are then fulfilled with exponential functions $\psi_l(M)=C_le^{-c_lM}$ for some $C_l,c_l>0$, $l=1,2$ and $V(Y')=N$.*

If the sub-gaussianity condition (1.20) is fulfilled for ξ', then $\psi_1(M)=e^{-\rho M}$. By Chebyshev's inequality, we see that the second relation in (A4) is fulfilled with function $\psi_2(M)=cM^{-2}$ and $V(Y')=N$ for any zero mean independent ξ'_i's with $\mathbb{E}\xi'^2_i=1$ and $\mathbb{E}[\xi'_i]^4\le C$.

Coming back to the problem of constructing a confidence ball of full coverage uniformly over Θ, let $\hat\theta$ and $\hat I$ be based on the sample Y and defined as before. We propose to mimic $\|\theta-\hat\theta\|^2$ by the de-biased quantity $\|Y'-\hat\theta\|^2-\sigma^2V(Y')$ plus additional $\sigma^2\sqrt{N}$-order term to control its oscillations, leading us to the following data dependent radius

$$\tilde R_M^2=\big(\|Y'-\hat\theta\|^2-\sigma^2V(Y')+2\sigma^2G_M\sqrt{N}\big)_+,\tag{5.27}$$

where $G_M=\sqrt{M(M+M_1)}$, $x_+=x\vee 0$ and the constant M_1 is from Theorem 5.1. The next theorem establishes the coverage and size properties of the confidence ball $B(\hat\theta,\tilde R_M)$.

Theorem 5.4. *Let Conditions* (A1), (A2) *and* (A4) *be fulfilled and* $\tilde{R}^2_M$ *be defined by* (5.27). *Then for any* $M \geq 0$

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, \tilde{R}_M)\big) \leq \psi_1(M/4) + \psi_2(M) + H_1 e^{-m_1 M},$$
$$\sup_{\theta\in\Theta} \mathbb{P}_\theta\big(\tilde{R}^2_M \geq g_M(\theta, N)\big) \leq \psi_1(M/4) + \psi_2(M) + 2H_1 e^{-m_1 M},$$

where $g_M(\theta, N) = M_1 r^2(\theta) + M\sigma^2 + 4\sigma^2 G_M \sqrt{N}$ *and the constants* H_1, m_1, M_1 *are defined in Theorem* 5.1.

By taking large enough M, we can ensure the coverage and size relations uniformly over the entire space Θ. Thus, the results of Theorem 5.4 are to be interpreted as the coverage and size relations in the optimality framework (1.14) with $\Theta_0 = \Theta_1 = \Theta$ and the effective radial rate $R(\theta) = \sqrt{g_M(\theta, N)} \asymp r(\theta) + \sigma N^{1/4}$ (for now disregarding the constants and the inflating factor M as we consider only the order of the radial rate). Since the both sets $\Theta_0 = \Theta_1 = \Theta = \mathbb{R}^N$ are the biggest possible, the deceptiveness phenomenon manifests itself only in the effective radial rate $R(\theta)$, which can be of a bigger order than the oracle rate $r(\theta)$ for $\theta \in \tilde{\Theta}$, where (for some $c > 0$)

$$\tilde{\Theta} = \tilde{\Theta}(c) = \{\theta \in \Theta : r^2(\theta) \leq c\sigma^2 N^{1/2}\}. \tag{5.28}$$

Equivalently, this can be seen as the optimality framework (1.14) with $\Theta_0 = \Theta = \mathbb{R}^N$, $\Theta_1 = \Theta\backslash\tilde{\Theta} = \mathbb{R}^N\backslash\tilde{\Theta}$ and the effective radial rate $R(\theta) \asymp r(\theta) + \sigma N^{1/4} \lesssim C r(\theta)$ is of the oracle rate order for $\theta \in \Theta_1$. Now the deceptiveness phenomenon manifests itself in the fact that $\Theta_1 = \Theta\backslash\tilde{\Theta}$, not the whole Θ.

In fact, the massiveness of the set $\tilde{\Theta}$ measures how much the deceptiveness phenomenon is present in particular models and structures. Loosely speaking, models and structures, where "good" estimation ($r^2(\theta) \lesssim \sigma^2 N^{1/2}$) is possible for "many" θ's ($\tilde{\Theta}$ is massive), suffer more from the deceptiveness phenomenon. For example, these are all models with sparsity structures in Section 5.5.7, as the set $\tilde{\Theta}$ is a substantial part of $\Theta = \mathbb{R}^N$ in those cases.

On the other hand, the deceptiveness phenomenon becomes effectively marginal for some "uninformative" models and structures, e.g., biclustering model (stochastic block model), dictionary learning (see Section 5.5), because in these cases the set $\tilde{\Theta}$ is a very "thin" subset of $\mathbb{R}^N$, informally described as set of *highly structured* parameters. In these cases, the extra term $\sigma N^{1/4}$ in the radial rate $R(\theta)$ does not increase its order as $\sigma N^{1/4} \lesssim r(\theta)$ for the "majority" of θ's: $\theta \in \Theta_1 = \Theta\backslash\tilde{\Theta}$. This means that, modulo the set $\tilde{\Theta}$ of highly structured parameters, there is no deceptiveness issue for those cases. Indeed, there is no payment in terms of removing deceptive parameters from the parameter space Θ in the coverage relation and the size relation holds uniformly over $\Theta_1 = \Theta\backslash\tilde{\Theta}$ which is "almost" the entire space Θ.

5.3. TECHNICAL LEMMAS

We provide a couple of technical lemmas used in the proofs of the main results. Recall that $\hat{\pi}(I|Y)$ is either $\tilde{\pi}(I|Y)$ defined by (5.10) or $\check{\pi}(I|Y) = 1\{I = \hat{I}\}$ defined by (5.14). In the latter case $\mathbb{E}_\theta \hat{\pi}(I|Y) = \mathbb{P}_\theta(\hat{I} = I)$. In what follows, denote $\hat{p}_I = \hat{\pi}(I|Y)$ for brevity.

Lemma 5.1. *Let Condition (A1) be fulfilled. Then for any $\theta \in \Theta$ and $I, I_0 \in \mathcal{I}$*

$$\mathbb{E}_\theta \hat{p}_I \le \big(\tfrac{\lambda_I}{\lambda_{I_0}}\big)^h \exp\big\{-\tfrac{1}{\sigma^2}(A_h\|\mathrm{P}_I^\perp\theta\|^2 - B_h\|\mathrm{P}_{I_0}^\perp\theta\|^2) + C_h\rho(s(I)) + D_h\rho(s(I_0))\big\},$$

where $h = \frac{\alpha}{4}$ and the constants $A_h = \frac{\alpha}{16}$, $B_h = \frac{3\alpha}{16}$ and $C_h = \frac{5}{8}$, $D_h = \frac{3+\alpha}{8}$.

If $\mathbb{L}_I \subseteq \mathbb{L}_{I_0}$, then

$$\mathbb{E}_\theta \hat{p}_I \le \big(\tfrac{\lambda_I}{\lambda_{I_0}}\big)^\alpha \exp\big\{-A_\alpha\sigma^{-2}\big(\|\mathrm{P}_I^\perp\theta\|^2 - \|\mathrm{P}_{I_0}^\perp\theta\|^2\big) + D_\alpha\rho(s(I_0))\big\},$$

where the constants $A_\alpha = \frac{\alpha}{3}$ and $D_\alpha = 1 + \frac{\alpha}{2}$.

If $\mathbb{L}_{I_0} \subseteq \mathbb{L}_I$, then

$$\mathbb{E}_\theta \hat{p}_I \le \big(\tfrac{\lambda_I}{\lambda_{I_0}}\big)^\alpha \exp\big\{B_\alpha\sigma^{-2}\big(\|\mathrm{P}_I\theta\|^2 - \|\mathrm{P}_{I_0}\theta\|^2\big) + C_\alpha\rho(s(I)) + D_\alpha\rho(s(I_0))\big\},$$

where the constants $B_\alpha = \alpha, C_\alpha = 1$ and $D_\alpha = \frac{\alpha}{2}$.

Proof. Recall that P_I is the projection onto $\mathbb{L}_I$. Since $\mathrm{P}_I - \mathrm{P}_{I_0} = \mathrm{P}_{I_0}^\perp - \mathrm{P}_I^\perp$, the bound

$$\begin{aligned} Y^T(\mathrm{P}_I - \mathrm{P}_{I_0})Y &= \theta^T(\mathrm{P}_I - \mathrm{P}_{I_0})\theta + 2\theta^T(\mathrm{P}_I - \mathrm{P}_{I_0})\sigma\xi + \sigma^2\xi^T(\mathrm{P}_I - \mathrm{P}_{I_0})\xi \\ &\le -\|\mathrm{P}_I^\perp\theta\|^2 + \|\mathrm{P}_{I_0}^\perp\theta\|^2 + 2|\sigma\theta^T(\mathrm{P}_I - \mathrm{P}_{I_0})\xi| + \sigma^2\|\mathrm{P}_I\xi\|^2 - \sigma^2\|\mathrm{P}_{I_0}\xi\|^2 \end{aligned} \tag{5.29}$$

holds for any $I, I_0 \in \mathcal{I}$. Using the relations $\mathrm{P}_I - \mathrm{P}_{I_0} = (\mathrm{P}_I - \mathrm{P}_{I_0})\mathrm{P}_{\mathbb{L}_I+\mathbb{L}_{I_0}}$, $\|\mathrm{P}_{\mathbb{L}_I+\mathbb{L}_{I_0}}x\|^2 \le \|\mathrm{P}_I x\|^2 + \|\mathrm{P}_{I_0}x\|^2$, $x \in \mathcal{Y}$, and the inequality $2ab \le a^2/4 + 4b^2$ (for any $a, b \in \mathbb{R}$), we derive

$$\begin{aligned} 2|\theta^T(\mathrm{P}_I - \mathrm{P}_{I_0})\sigma\xi| &= 2|\theta^T(\mathrm{P}_I - \mathrm{P}_{I_0})\mathrm{P}_{\mathbb{L}_I+\mathbb{L}_{I_0}}\sigma\xi| \\ &\le 2\|\theta^T(\mathrm{P}_I - \mathrm{P}_{I_0})\|\|\sigma\mathrm{P}_{\mathbb{L}_I+\mathbb{L}_{I_0}}\xi\| \le \tfrac{1}{4}\|(\mathrm{P}_I - \mathrm{P}_{I_0})\theta\|^2 + 4\sigma^2\|\mathrm{P}_{\mathbb{L}_I+\mathbb{L}_{I_0}}\xi\|^2 \\ &= \tfrac{1}{4}\|(\mathrm{P}_{I_0}^\perp - \mathrm{P}_I^\perp)\theta\|^2 + 4\sigma^2\|\mathrm{P}_{\mathbb{L}_I+\mathbb{L}_{I_0}}\xi\|^2 \\ &\le \tfrac{1}{2}\|\mathrm{P}_I^\perp\theta\|^2 + \tfrac{1}{2}\|\mathrm{P}_{I_0}^\perp\theta\|^2 + 4\sigma^2\|\mathrm{P}_I\xi\|^2 + 4\sigma^2\|\mathrm{P}_{I_0}\xi\|^2. \end{aligned}$$

The last bound and (5.29) imply that

$$Y^T(\mathrm{P}_I - \mathrm{P}_{I_0})Y \le -\tfrac{1}{2}\|\mathrm{P}_I^\perp\theta\|^2 + \tfrac{3}{2}\|\mathrm{P}_{I_0}^\perp\theta\|^2 + 5\sigma^2\|\mathrm{P}_I\xi\|^2 + 3\sigma^2\|\mathrm{P}_{I_0}\xi\|^2. \tag{5.30}$$

In case $\hat{p}_I = \hat{\pi}(I|Y) = \check{\pi}(I|Y) = 1\{\hat{I} = I\}$, (5.10), the definition (5.12) of $\hat{I}$ and the Markov inequality imply that, for any $I, I_0 \in \mathcal{I}$ and any $h \ge 0$,

$$\mathbb{E}_\theta \hat{p}_I = \mathbb{P}_\theta(\hat{I} = I) \le \mathbb{P}_\theta\Big(\frac{\tilde{\pi}(I|Y)}{\tilde{\pi}(I_0|Y)} \ge 1\Big) \le \mathbb{E}_\theta\Big[\frac{\tilde{\pi}(I|Y)}{\tilde{\pi}(I_0|Y)}\Big]^h. \tag{5.31}$$

In case $\hat{p}_I = \hat{\pi}(I|Y) = \tilde{\pi}(I|Y)$, (5.10) implies (5.31) for any $I, I_0 \in \mathcal{I}$, $h \in [0,1]$.

Combining (5.30) and (5.31), we derive for any $I, I_0 \in \mathcal{I}$ and any $h \in [0,1]$,

$$\begin{aligned} \mathbb{E}_\theta \hat{p}_I &\le \mathbb{E}_\theta\Big[\frac{\lambda_I \exp\{-\frac{1}{2\sigma^2}[\|Y - \mathrm{P}_I Y\|^2 + \sigma^2\dim(\mathbb{L}_I)]\}}{\lambda_{I_0}\exp\{-\frac{1}{2\sigma^2}[\|Y - \mathrm{P}_{I_0}Y\|^2 + \sigma^2\dim(\mathbb{L}_{I_0})]\}}\Big]^h \\ &= \big(\tfrac{\lambda_I}{\lambda_{I_0}}\big)^h \mathbb{E}_\theta \exp\big\{\tfrac{h}{2\sigma^2}\big(Y^T(\mathrm{P}_I - \mathrm{P}_{I_0})Y + \sigma^2\dim(\mathbb{L}_{I_0}) - \sigma^2\dim(\mathbb{L}_I)\big)\big\} \\ &\le \big(\tfrac{\lambda_I}{\lambda_{I_0}}\big)^h e^{-\frac{h}{4\sigma^2}\|\mathrm{P}_I^\perp\theta\|^2 + \frac{3h}{4\sigma^2}\|\mathrm{P}_{I_0}^\perp\theta\|^2 + \frac{h}{2}\rho(s(I_0))}\mathbb{E}e^{\frac{h}{2}(5\|\mathrm{P}_I\xi\|^2 + 3\|\mathrm{P}_{I_0}\xi\|^2)}. \end{aligned} \tag{5.32}$$

The lemma follows for $h = \frac{\alpha}{4}$ from the last display and the relation

$$\begin{aligned}\mathbb{E}_\theta \exp\{\tfrac{5\alpha}{8}\|\mathrm{P}_I\xi\|^2 + \tfrac{3\alpha}{8}\|\mathrm{P}_{I_0}\xi\|^2\} &\le \big[\mathbb{E}_\theta e^{\alpha\|\mathrm{P}_I\xi\|^2}\big]^{\frac{5}{8}}\big[\mathbb{E}_\theta e^{\alpha\|\mathrm{P}_{I_0}\xi\|^2}\big]^{\frac{3}{8}}\\ &\le \exp\{\tfrac{5}{8}\rho(s(I)) + \tfrac{3}{8}\rho(s(I_0))\},\end{aligned}$$

which is in turn obtained by using the Hölder inequality and Condition (A1).

In case $\mathbb{L}_I \subseteq \mathbb{L}_{I_0}$, take $h = \alpha$ in (5.32) and, instead of (5.30) use $Y^T(\mathrm{P}_I - \mathrm{P}_{I_0})Y = -\|\mathrm{P}_{\mathbb{L}_I^\perp\cap\mathbb{L}_{I_0}}Y\|^2 \le -\frac{2}{3}\|\mathrm{P}_{\mathbb{L}_I^\perp\cap\mathbb{L}_{I_0}}\theta\|^2 + 2\sigma^2\|\mathrm{P}_{\mathbb{L}_I^\perp\cap\mathbb{L}_{I_0}}\xi\|^2 \le \frac{2}{3}\big(\|\mathrm{P}_I\theta\|^2 - \|\mathrm{P}_{I_0}\theta\|^2\big) + 2\sigma^2\|\mathrm{P}_{I_0}\xi\|^2 = -\frac{2}{3}\|\mathrm{P}_I^\perp\theta\|^2 + \frac{2}{3}\|\mathrm{P}_{I_0}^\perp\theta\|^2 + 2\sigma^2\|\mathrm{P}_{I_0}\xi\|^2$ as $(a+b)^2 \ge 2a^2/3 - 2b^2$ and $\mathrm{P}_{I_0} - \mathrm{P}_I = \mathrm{P}_{\mathbb{L}_I^\perp\cap\mathbb{L}_{I_0}}$.

In case $\mathbb{L}_{I_0} \subseteq \mathbb{L}_I$, take $h = \alpha$ in (5.32) and, instead of (5.30) use $Y^T(\mathrm{P}_I - \mathrm{P}_{I_0})Y = \|\mathrm{P}_{\mathbb{L}_I\cap\mathbb{L}_{I_0}^\perp}Y\|^2 \le 2\|\mathrm{P}_{\mathbb{L}_I\cap\mathbb{L}_{I_0}^\perp}\theta\|^2 + 2\sigma^2\|\mathrm{P}_{\mathbb{L}_I\cap\mathbb{L}_{I_0}^\perp}\xi\|^2 \le 2\big(\|\mathrm{P}_I\theta\|^2 - \|\mathrm{P}_{I_0}\theta\|^2\big) + 2\sigma^2\|\mathrm{P}_I\xi\|^2 = -2\|\mathrm{P}_I^\perp\theta\|^2 + 2\|\mathrm{P}_{I_0}^\perp\theta\|^2 + 2\sigma^2\|\mathrm{P}_I\xi\|^2$ as $(a+b)^2 \le 2a^2 + 2b^2$ and $\mathrm{P}_I - \mathrm{P}_{I_0} = \mathrm{P}_{\mathbb{L}_I\cap\mathbb{L}_{I_0}^\perp}$. □

Note that above lemma holds for any $I_0 \in \mathcal{I}$. By taking $I_0 = I_o$ defined by (5.15), we derive the next lemma.

Lemma 5.2. *Let Condition (A1) be fulfilled. Then there exist positive constants $c_1 = c_1(\varkappa) > 2\nu, c_2$ and $c_3 = c_3(\varkappa)$ such that for any $\theta \in \Theta$*

$$\mathbb{E}_\theta \hat{p}_I \le \exp\big\{-c_1\rho(s(I)) - c_2\sigma^{-2}\big[r^2(I,\theta) - c_3 r^2(\theta)\big]\big\}.$$

Proof. With constants h, A_h, B_h, C_h, D_h defined in Lemma 5.1, define the constant $c_1 = c_1(\varkappa) = h\varkappa - C_h - A_h = \frac{\varkappa\alpha}{4} - \frac{5}{8} - \frac{\alpha}{16} > 2\nu$ as $\varkappa > \bar{\varkappa}$ by (5.4). The definition (5.3) of λ_I entails that

$$\big(\lambda_I/\lambda_{I_o}\big)^h = \exp\big\{h\varkappa\rho(s(I_o)) - \big(c_1 + C_h + A_h\big)\rho(s(I))\big\}.$$

Combining the last relation with Lemma 5.1 (for $I_0 = I_o$), we derive that

$$\begin{aligned}\mathbb{E}_\theta \hat{p}_I &\le \exp\big\{-c_1\rho(s(I)) - \sigma^{-2}\big[A_h r^2(I,\theta) - \max\{B_h, D_h + h\varkappa\}r^2(\theta)\big]\big\}\\ &= \exp\big\{-c_1\rho(s(I)) - c_2\sigma^{-2}\big[r^2(I,\theta) - c_3 r^2(\theta)\big]\big\},\end{aligned}$$

which completes the proof with the constants $c_1 = c_1(\varkappa) = \frac{\varkappa\alpha}{4} - \frac{5}{8} - \frac{\alpha}{16}$, $c_2 = A_h = \frac{\alpha}{16}$ and $c_3 = c_3(\varkappa) = A_h^{-1}\max\{B_h, D_h + h\varkappa\} = \frac{16}{\alpha}\max\{\frac{3\alpha}{16}, \frac{3+\alpha}{8} + \frac{\alpha\varkappa}{4}\} = \max\{3, \frac{6}{\alpha} + 2 + 4\varkappa\} = \frac{6}{\alpha} + 2 + 4\varkappa$ because $\varkappa > \bar{\varkappa} \ge 1$ by (5.4). □

5.4. PROOFS OF THE THEOREMS

Here we gather the proofs of all the theorems. By C_1, C_2 etc., we denote constants which are different in different proofs.

Proof. of Theorem 5.1 Recall the constants c_1, c_2, c_3 defined in the proof of Lemma 5.2 and the notation $\hat{p}_I = \hat{\pi}(I|Y)$. For any $\theta \in \Theta$, $M \ge 0$ and some constant M_0 to be chosen later, denote $\Delta_M = \Delta_M(\theta) = M_0 r^2(\theta) + M\sigma^2$. Next, introduce the set $\mathcal{O}_M = \mathcal{O}_M(\theta) = \{I \in$

$\mathcal{I}: r^2(I,\theta) \le c_3 r^2(\theta) + C_1 M\sigma^2\}$ and the events $A_M(I) = \{\alpha\|\mathrm{P}_I\xi\|^2 \le (\nu+1)\rho(s(I)) + C_2 M\}$, $I \in \mathcal{I}$, where constants $C_1, C_2 > 0$ are to be chosen later. We have

$$\begin{aligned}
\hat\pi(\|\vartheta-\theta\|^2 \ge \Delta_M|Y) &= \sum_{I\in\mathcal{I}} \hat\pi_I(\|\vartheta-\theta\|^2 \ge \Delta_M|Y)\hat p_I \\
&\le \sum_{I\in\mathcal{I}} \hat p_I 1_{A_M^c(I)} + \sum_{I\in\mathcal{O}_M^c} \hat p_I + \sum_{I\in\mathcal{O}_M} \hat\pi_I\big(\|\vartheta-\theta\|^2 \ge \Delta_M|Y\big)\hat p_I 1_{A_M(I)} \\
&= T_1 + T_2 + T_3. \qquad (5.33)
\end{aligned}$$

Now we need to bound the quantities $\mathbb{E}_\theta T_1$, $\mathbb{E}_\theta T_2$ and $\mathbb{E}_\theta T_3$.

By using the Markov inequality and Condition (A1), we have

$$\mathbb{P}_\theta(A_M^c(I)) = \mathbb{P}_\theta\big(e^{\alpha\|\mathrm{P}_I\xi\|^2} > e^{(\nu+1)\rho(s(I))+C_2M}\big) \le e^{-\nu\rho(s(I))-C_2M}.$$

The last relation and Condition (A2) yield the bound for $\mathbb{E}_\theta T_1$:

$$\mathbb{E}_\theta T_1 \le \sum_{I\in\mathcal{I}} \mathbb{P}_\theta\big(A_M^c(I)\big) \le \sum_{I\in\mathcal{I}} \exp\{-\nu\rho(s(I)) - C_2M\} \le C_\nu e^{-C_2M}. \qquad (5.34)$$

If $I \in \mathcal{O}_M^c$, then $r^2(I,\theta) > c_3 r^2(\theta) + C_1 M\sigma^2$. Using this, Lemma 5.2 and the fact that $\sum_{I\in\mathcal{I}} e^{-c_1\rho(s(I))} \le C_\nu$ (in view of Condition (A2) and because $c_1 > 2\nu$), we bound $\mathbb{E}_\theta T_2$ as follows:

$$\begin{aligned}
\mathbb{E}_\theta T_2 = \sum_{I\in\mathcal{O}_M^c} \mathbb{E}_\theta \hat p_I &\le \sum_{I\in\mathcal{O}_M^c} \exp\big\{-c_1\rho(s(I)) - c_2\sigma^{-2}\big[r^2(I,\theta) - c_3 r^2(\theta)\big]\big\} \\
&\le \sum_{I\in\mathcal{I}} \exp\big\{-c_1\rho(s(I)) - c_2C_1M\big\} \le C_\nu \exp\big\{-c_2C_1M\big\}. \qquad (5.35)
\end{aligned}$$

It remains to establish the last bound for $\mathbb{E}_\theta T_3$. For $I \in \mathcal{O}_M$, we have that

$$A_M(I) \subseteq \Big\{\|\theta-\mathrm{P}_I\theta\|^2 + \sigma^2\|\mathrm{P}_I\xi\|^2 \le c_3 r^2(\theta) + \tfrac{\nu+1}{\alpha}\sigma^2\rho(s(I)) + (C_1 + \tfrac{C_2}{\alpha})M\sigma^2\Big\}.$$

Recall that, in view of (5.9), $\hat\pi_I(\vartheta|Y) = \mathrm{N}(\mathrm{P}_I Y, (1-e^{-1})\sigma^2\mathrm{P}_I)$. Let $\mathbb{P}_Z$ be the measure of $Z = (Z_1,\dots,Z_N)$, with $Z_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. In Remark 1.6, we established $\mathbb{E}e^{0.4\|\mathrm{P}_I Z\|^2} \le e^{d_I}$, which implies $\mathbb{P}_Z\big(\|\mathrm{P}_I Z\|^2 \ge \frac{5}{2}d_I + M\big) \le e^{-2M/5}$, $I \in \mathcal{I}$. Using this, the last display and the fact that $\frac{r^2(\theta)}{\sigma^2} \ge c_3^{-1}(\rho(s(I)) - C_1 M)$ for $I \in \mathcal{O}_M$, we obtain that, for any $I \in \mathcal{O}_M$,

$$\begin{aligned}
&\hat\pi_I\big(\|\vartheta-\theta\|^2 \ge \Delta_M|Y\big)1_{A_M(I)} \\
&\quad= \mathbb{P}_Z(\|\mathrm{P}_I Y + (1-e^{-1})^{1/2}\sigma\mathrm{P}_I Z - \theta\|^2 \ge \Delta_M)1_{A_M(I)} \\
&\quad\le \mathbb{P}_Z\big(2\sigma^2\|\mathrm{P}_I Z\|^2 + 2\|\mathrm{P}_I Y - \theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2\big)1_{A_M(I)} \\
&\quad= \mathbb{P}_Z\big(\sigma^2\|\mathrm{P}_I Z\|^2 + \|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\|\mathrm{P}_I\xi\|^2 \ge \tfrac{M_0}{2}r^2(\theta) + \tfrac{M\sigma^2}{2}\big)1_{A_M(I)} \\
&\quad\le \mathbb{P}_Z\Big(\|\mathrm{P}_I Z\|^2 \ge (\tfrac{M_0}{2} - c_3)\tfrac{r^2(\theta)}{\sigma^2} - \tfrac{\nu+1}{\alpha}\rho(s(I)) + \tfrac{M}{2} - (C_1 + \tfrac{C_2}{\alpha})M\Big) \\
&\quad\le \mathbb{P}_Z\Big(\|\mathrm{P}_I Z\|^2 \ge \big(\tfrac{M_0}{2c_3} - \tfrac{\nu+\alpha+1}{\alpha}\big)\rho(s(I)) + \tfrac{M}{2} - \tfrac{C_2}{\alpha}M - \tfrac{M_0C_1}{2c_3}M\Big) \\
&\quad= \mathbb{P}_Z\Big(\|\mathrm{P}_I Z\|^2 \ge \tfrac{5}{2}\rho(s(I)) + \tfrac{M}{4}\Big) \le e^{-M/10},
\end{aligned}$$

where we have chosen $M_0 = \frac{c_3(2\nu+7\alpha+2)}{\alpha}$, $C_1 = \frac{\alpha}{4(2\nu+7\alpha+2)}$ and $C_2 = \frac{\alpha}{8}$ (so that $\frac{M_0}{2c_3} - \frac{\nu+\alpha+1}{\alpha} = \frac{5}{2}$, $\frac{C_2}{\alpha} = \frac{1}{8}$, $\frac{M_0 C_1}{2c_3} = \frac{1}{8}$). Thus we have derived

$$\mathbb{E}_\theta T_3 = \mathbb{E}_\theta \sum_{I\in\mathcal{O}_M} 1_{A_M(I)} \hat{\pi}_I\big(\|\vartheta-\theta\|^2 \ge \Delta_M | Y\big)\hat{p}_I \le \mathbb{E}_\theta \sum_{I\in\mathcal{I}} e^{-\frac{M}{10}}\hat{p}_I \le e^{-M/10}.$$

This completes the proof of the first assertion since, in view of (5.33), (5.34), (5.35) and the last display, we established the claim (5.17): $\mathbb{E}_\theta\hat{\pi}\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2|Y\big) \le \mathbb{E}_\theta(T_1+T_2+T_3) \le H_0 e^{-m_0 M}$, with the constants $M_0 = \frac{c_3(2\nu+7\alpha+2)}{\alpha}$, $H_0 = 1+2C_\nu$ and $m_0 = \min\{C_2, c_2C_1, 1/10\}$.

The proof of the assertion (5.18) proceeds along similar lines. Introduce the set $\mathcal{J}_M = \mathcal{J}_M(\theta) = \{I\in\mathcal{I}: r^2(I,\theta) \le 2c_3 r^2(\theta) + C_3 M\sigma^2\}$ and the events $B_M(I) = \big\{\alpha\|\mathrm{P}_I\xi\|^2 \le 2(\nu+1)\rho(s(I)) + C_4 M\big\}$, $I\in\mathcal{I}$, where constants $C_3, C_4 > 0$ are to be chosen later.

If $M\in[0,1]$, the claim (ii) holds for $H_1 = e^{m_1}$. Let $M \ge 1$. Denote for brevity $R_I^2 = R_I^2(\theta,Y) = \|\theta - \mathrm{P}_I Y\|^2 = \|\theta-\mathrm{P}_I\theta\|^2 + \sigma^2\|\mathrm{P}_I\xi\|^2$, $\Delta'_M = \Delta'_M(\theta) = M_1 r^2(\theta) + M\sigma^2$ and $\hat{p}_I = \hat{\pi}(I|Y)$, where $M_1 > 0$ is to be chosen later. Applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{P}_\theta(\|\hat{\theta}-\theta\|^2 \ge \Delta'_M) \le \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{I}} R_I^2\hat{p}_I \ge \Delta'_M\Big)\\
&\le \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}_M} R_I^2\hat{p}_I(1_{B_M(I)} + 1_{B_M^c(I)}) + \sum_{I\in\mathcal{J}_M^c} R_I^2\hat{p}_I \ge \Delta'_M\Big)\\
&\le \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}_M} R_I^2\hat{p}_I 1_{B_M(I)} \ge \tfrac{\Delta'_M}{3}\Big) + \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}_M} R_I^2\hat{p}_I 1_{B_M^c(I)} \ge \tfrac{\Delta'_M}{3}\Big)\\
&\quad + \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}_M^c} R_I^2\hat{p}_I \ge \tfrac{\Delta'_M}{3}\Big) = \bar{T}_1 + \bar{T}_2 + \bar{T}_3. \qquad (5.36)
\end{aligned}$$

Let us evaluate $\bar{T}_1$. For any $I\in\mathcal{J}_M$, under $B_M(I)$, we have that $R_I^2 = \|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\|\mathrm{P}_I\xi\|^2 \le \|\theta-\mathrm{P}_I\theta\|^2 + \frac{2(\nu+1)}{\alpha}\sigma^2\rho(s(I)) + \frac{C_4}{\alpha}M\sigma^2 \le \frac{2(\nu+1)}{\alpha}r^2(I,\theta) + \frac{C_4}{\alpha}M\sigma^2 \le \frac{4c_3(\nu+1)}{\alpha}r^2(\theta) + \frac{2C_3(\nu+1)+C_4}{\alpha}M\sigma^2$. Using this, we derive

$$\begin{aligned}
\bar{T}_1 &= \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}_M} R_I^2\hat{p}_I 1_{B_M(I)} \ge \tfrac{\Delta'_M}{3}\Big)\\
&\le \mathbb{P}_\theta\Big(\tfrac{4c_3(\nu+1)}{\alpha}r^2(\theta) + \tfrac{2C_3(\nu+1)+C_4}{\alpha}M\sigma^2 \ge \tfrac{\Delta'_M}{3}\Big) = 0, \qquad (5.37)
\end{aligned}$$

as $\frac{4c_3(\nu+1)}{\alpha} = \frac{M_1}{3}$ and $\frac{2C_3(\nu+1)+C_4}{\alpha} < \frac{1}{3}$ because we choose $M_1 = \frac{12c_3(\nu+1)}{\alpha}$, $C_3 = \frac{\alpha}{12(\nu+1)}$ and $C_4 = \frac{\alpha}{7}$.

Next, we evaluate $\bar{T}_2$. By Condition (A1) and the Markov inequality,

$$\mathbb{P}_\theta(B_M^c(I)) = \mathbb{P}_\theta\big(\alpha\|\mathrm{P}_I\xi\|^2 > (2\nu+2)\rho(s(I)) + C_4M\big) \le e^{-(2\nu+1)\rho(s(I))-C_4M}.$$

It follows from (1.22) with $t=\frac{1}{2}$ that $\big[\mathbb{E}_\theta\|\mathrm{P}_I\xi\|^4\big]^{1/2} \le \frac{2}{\alpha}\exp\{\rho(s(I))/2\}$ for any $I\in\mathcal{I}$. By Condition (A2), $\sum_{I\in\mathcal{I}}\exp\{-\nu\rho(s(I))\} \le C_\nu$. Besides, for any $I\in\mathcal{J}_M$, $\|\theta-\mathrm{P}_I\theta\|^2/\Delta'_M \le$

$(2c_3 r^2(\theta) + C_3 M\sigma^2)/(M_1 r^2(\theta) + M\sigma^2) \le \frac{2c_3}{M_1} + C_3$ and $\Delta'_M \ge M\sigma^2 \ge \sigma^2$ (as $M \ge 1$). Collecting all the derived relations for evaluating T_2 and using the Markov and Cauchy-Schwarz inequalities, we obtain

$$\begin{aligned}
\bar{T}_2 &= \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}_M} R_I^2 \hat{p}_I 1_{B^c_M(I)} \ge \Delta'_M/3\Big) \\
&\le \frac{\mathbb{E}_\theta \sum_{I\in\mathcal{J}_M} \big(\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\|\mathrm{P}_I\xi\|^2\big)\hat{p}_I 1_{B^c_M(I)}}{\Delta'_M/3} \\
&\le \frac{\sum_{I\in\mathcal{J}_M}\|\theta - \mathrm{P}_I\theta\|^2 \mathbb{P}_\theta(B^c_M(I))}{\Delta'_M/3} + \frac{\sigma^2 \sum_{I\in\mathcal{J}_M}\big[\mathbb{E}_\theta\|\mathrm{P}_I\xi\|^4\big]^{1/2}\big[\mathbb{P}_\theta(B^c_M(I))\big]^{1/2}}{\Delta'_M/3} \\
&\le 3\big(\tfrac{2c_3}{M_1} + C_3\big)e^{-C_4 M} \sum_{I\in\mathcal{J}_M} e^{-(2\nu+1)\rho(s(I))} + \tfrac{6}{\alpha} e^{-C_4 M/2} \sum_{I\in\mathcal{J}_M} e^{-\nu\rho(s(I))} \\
&\le 3C_\nu\big(\tfrac{2c_3}{M_1} + C_3\big)e^{-C_4 M} + \tfrac{6C_\nu}{\alpha} e^{-C_4 M/2}. \qquad (5.38)
\end{aligned}$$

It remains to bound T_3. Applying first the Markov inequality and then the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\bar{T}_3 = \mathbb{P}_\theta\Big(\sum_{I\in\mathcal{J}^c_M} R_I^2 \hat{p}_I \ge \Delta'_M/3\Big) &\le \frac{\sum_{I\in\mathcal{J}^c_M}\|\theta - \mathrm{P}_I\theta\|^2 \mathbb{E}_\theta \hat{p}_I}{\Delta'_M/3} \\
&+ \frac{\sigma^2 \sum_{I\in\mathcal{J}^c_M}\big(\mathbb{E}_\theta\|\mathrm{P}_I\xi\|^4\big)^{1/2}\big[\mathbb{E}_\theta \hat{p}_I\big]^{1/2}}{\Delta'_M/3} = \bar{T}_{31} + \bar{T}_{32}. \qquad (5.39)
\end{aligned}$$

For each $I \in \mathcal{J}^c_M$, we have $c_3 r^2(\theta) \le \frac{1}{2} r^2(I,\theta) - \frac{C_3}{2} M\sigma^2$, yielding the bound

$$\frac{c_2}{2\sigma^2}\big(r^2(I,\theta) - c_3 r^2(\theta)\big) \ge \frac{c_2}{4\sigma^2} r^2(I,\theta) + \frac{c_2 C_3}{4} M.$$

The last relation and Lemma 5.2 entail that, for each $I \in \mathcal{J}^c_M$,

$$\big[\mathbb{E}_\theta \hat{p}_I\big]^{1/2} \le \exp\Big\{-\frac{c_1}{2}\rho(s(I)) - \frac{c_2}{4\sigma^2} r^2(I,\theta) - \frac{c_2 C_3}{4} M\Big\}. \qquad (5.40)$$

Since $M \ge 1$, $\Delta'_M \ge M\sigma^2 \ge \sigma^2$. Using this, the relation (5.40), the facts that $\max_{x\ge 0}\{xe^{-cx}\} \le (ce)^{-1}$ (for any $c > 0$) and $\sum_{I\in\mathcal{I}} e^{-c_1\rho(s(I))} \le C_\nu$ (in view of Condition (A2) as $c_1 > 2\nu$), we bound the term T_{31} as follows:

$$\begin{aligned}
\bar{T}_{31} &= \frac{\sum_{I\in\mathcal{J}^c_M}\|\theta - \mathrm{P}_I\theta\|^2 \mathbb{E}_\theta \hat{p}_I}{\Delta'_M/3} \\
&\le 3 \sum_{I\in\mathcal{J}^c_M} \frac{r^2(I,\theta)}{\sigma^2} \exp\big\{-c_1\rho(s(I)) - \tfrac{c_2}{2\sigma^2} r^2(I,\theta_0) - \tfrac{c_2 C_3}{2} M\big\} \\
&\le \frac{6C_\nu}{c_2 e} e^{-c_2 C_3 M/2}. \qquad (5.41)
\end{aligned}$$

Using (1.22) with $t_0 = \min\{1/2, c_2/4\}$, we have that $\big[\mathbb{E}_\theta\|\mathrm{P}_I\xi\|^4\big]^{1/2} \le \frac{1}{\alpha t_0}\exp\{\frac{c_2}{4}\rho(s(I))\}$. Besides, $\Delta'_M \ge \sigma^2$, $r^2(I,\theta) \ge \sigma^2\rho(s(I))$ and, as $c_1 > 2\nu$, $\sum_{I\in\mathcal{I}} e^{-c_1\rho(s(I))/2} \le C_\nu$. Piecing all

these together with (5.40), we obtain

$$\begin{aligned}\bar{T}_{32} &= \frac{\sigma^2 \sum_{I\in\mathcal{J}_M^c} \left(\mathbb{E}_\theta \|\mathrm{P}_I \xi\|^4\right)^{1/2} \left[\mathbb{E}_\theta \hat{p}_I\right]^{1/2}}{\Delta'_M/3} \\ &\le \frac{3e^{-c_2C_3M/4}}{\alpha\, t_0} \sum_{I\in\mathcal{J}_M^c} \exp\left\{-\tfrac{c_1}{2}\rho(s(I)) + \tfrac{c_2}{4}\rho(s(I)) - \tfrac{c_2}{4\sigma^2} r^2(I,\theta)\right\} \\ &\le \frac{3e^{-c_2C_3M/4}}{\alpha\, t_0} \sum_{I\in\mathcal{I}} \exp\left\{-\tfrac{c_1}{2}\rho(s(I))\right\} \le \frac{3C_\nu}{\alpha\, t_0}\, e^{-c_2C_3M/4}.\end{aligned}$$

Combining (5.36), (5.37), (5.38), (5.39), (5.41) and the last relation finishes the proof of claim (5.18) with the constants $M_1 = \frac{12c_3(\nu+1)}{\alpha}$, $H_1 = \max\{C_\nu\big[3(\frac{2c_3}{M_1} + C_3) + \frac{6}{\alpha} + \frac{6}{c_2 e} + \frac{3}{\alpha\min\{1/2, c_2/4\}}\big], e^{m_1}\}$ and $m_1 = \min\{\frac{C_4}{2}, \frac{c_2C_3}{4}\}$.

□

Proof. of Theorem 5.2 First we prove (i). Denote $\mathcal{G}_1 = \mathcal{G}_1(\theta, M) = \{I \in \mathcal{I} : r^2(I,\theta) \ge c_3 r^2(\theta) + M\sigma^2\}$, where the constants $c_1 > 2\nu$, c_2, c_3 are defined in Lemma 5.2. Applying Lemma 5.2 and Condition (A2), we obtain

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{G}_1 \big| Y\big) = \sum_{I\in\mathcal{G}_1} \mathbb{E}_\theta \hat{\pi}(I|Y) \le e^{-c_2M} \sum_{I\in\mathcal{I}} e^{-c_1\rho(s(I))} \le C_\nu e^{-c_2M},$$

which completes the proof of (i).

Now we prove (ii). By Condition (A3), for any $I, I_1 \in \mathcal{I}$ there exists $I' = I'(I, I_1) \in \mathcal{I}$ such that $(\mathbb{L}_I \cup \mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'}$. Fix $I_1 \in \mathcal{I}$ and define $\mathcal{G}_2(M, I_1) = \{I \in \mathcal{I} : \theta^T[\mathrm{P}_{I'} - \mathrm{P}_I]\theta \ge \bar{\tau}\rho(s(I'))\sigma^2 + M\sigma^2\}$, where $\bar{\tau}$ is defined by (5.20).

As $\mathbb{L}_I \subseteq \mathbb{L}_{I'}$, by using (5.3) and applying Lemma 5.1 with $h = \alpha$ and $I_0 = I'$, we obtain that, for each $I \in \mathcal{G}_2(M, I_1)$,

$$\begin{aligned}\mathbb{E}_\theta \hat{p}_I &\le \left(\tfrac{\lambda_I}{\lambda_{I'}}\right)^\alpha \exp\left\{-\tfrac{\alpha}{3\sigma^2}\left[\theta^T(\mathrm{P}_{I'} - \mathrm{P}_I)\theta\right] + (1 + \tfrac{\alpha}{2})\rho(s(I'))\right\} \\ &= \exp\left\{-\varkappa\alpha\rho(s(I)) - \tfrac{\alpha}{3\sigma^2}\left[\theta^T(\mathrm{P}_{I'} - \mathrm{P}_I)\theta\right] + (1 + \tfrac{\alpha}{2} + \varkappa\alpha)\rho(s(I'))\right\} \\ &\le \exp\left\{-\varkappa\alpha\rho(s(I)) - \left[\tfrac{\alpha\bar{\tau}}{3} - (1 + \tfrac{\alpha}{2} + \varkappa\alpha)\right]\rho(s(I')) - \tfrac{\alpha}{3}M\right\} \\ &= e^{-\varkappa\alpha\rho(s(I)) - \frac{\alpha}{3}M}.\end{aligned}$$

Since $\varkappa \ge \alpha^{-1}\nu$, by Condition (A2) we have that $\sum_{I\in\mathcal{I}} e^{-\varkappa\alpha\rho(s(I))} \le C_\nu$. This relation and the last display imply that, with $m'_0 = \alpha/3$,

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{G}_2(M, I_1) | Y\big) = \sum_{I\in\mathcal{G}_2(M,I_1)} \mathbb{E}_\theta \hat{p}_I \le C_\nu \exp\left\{-m'_0 M\right\}. \tag{5.42}$$

Now take $I_1 = I_*$ defined by (5.21). By Condition (A3) there exists $I'(I, I_*) \in \mathcal{I}$ such that $(\mathbb{L}_I \cup \mathbb{L}_{I_*}) \subseteq \mathbb{L}_{I'}$ and $\rho(s(I')) \le \rho(s(I)) + \rho(s(I_*))$. If $\rho(s(I)) \le \delta\rho(s(I_*)) - M$, then $\rho(s(I')) \le \rho(s(I)) + \rho(s(I_*)) \le (1+\delta)\rho(s(I_*)) - M$. Hence, $\rho(s(I_*)) \ge \frac{1}{1+\delta}\rho(s(I')) + \frac{M}{1+\delta}$ and $\mathrm{P}_{I'} \ge \mathrm{P}_{I_*}$, which, together with the definition (5.15) of the τ-oracle, imply

$$\begin{aligned}\theta^T[\mathrm{P}_{I'} - \mathrm{P}_I]\theta &\ge \theta^T[\mathrm{P}_{I_*} - \mathrm{P}_I]\theta \ge \tau_0\sigma^2[\rho(s(I_*)) - \rho(s(I))] \\ &\ge \tau_0\sigma^2(1-\delta)\rho(s(I_*)) + \tau_0 M\sigma^2 \ge \tau'\sigma^2\rho(s(I')) + \tau_0 M\sigma^2,\end{aligned}$$

5

where $\tau' \triangleq \frac{1-\delta}{1+\delta}\tau_0 > \bar{\tau}$ by the condition of the theorem. It follows that $\{I \in \mathcal{I} : \rho(s(I)) \le \delta\rho(s(I_*)) - M\} \subseteq \mathcal{G}_2(\tau_0 M, I_*)$. Thus, we obtain

$$\mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : \rho(s(I)) \le \delta\rho(s(I_*)) - M \big| Y\big) \le \mathbb{E}_\theta \hat{\pi}\big(\mathcal{G}_2(\tau_0 M, I_*)|Y\big).$$

The last relation and (5.42) imply claim (ii) with $m_1' = \tau_0 m_0' = \tau_0 \alpha/3$.

Finally, we prove (iii). Condition (A3') implies that $\mathbb{L}_{I'} = \mathbb{L}_{I_o} + \mathbb{L}_I = \mathbb{L}_{I_o} \oplus (\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp)$. If the inequality $\sigma^2\rho(s(I)) < \|\mathrm{P}_{\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp}\theta\|^2$ would hold, then

$$\begin{aligned} r^2(I',\theta) &= \|\theta - \mathrm{P}_{I'}\theta\|^2 + \sigma^2\rho(s(I')) \\ &\le \|\theta - (\mathrm{P}_{I_o} + \mathrm{P}_{\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp})\theta\|^2 + \sigma^2(\rho(s(I_o)) + \rho(s(I)) \\ &< \|\mathrm{P}_{\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp}\theta\|^2 + \|\theta - (\mathrm{P}_{I_o} + \mathrm{P}_{\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp})\theta\|^2 + \sigma^2\rho(s(I_o)) \\ &= \|\theta - \mathrm{P}_{I_o}\theta\|^2 + \sigma^2\rho(s(I_o)) = r^2(\theta), \end{aligned}$$

which contradicts the definition of the oracle. Hence, $\|\mathrm{P}_{\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp}\theta\|^2 \le \sigma^2\rho(s(I))$.

Take $I_0 \in \mathcal{I}$ such that $\mathbb{L}_{I_0} = \mathbb{L}_I \cap \mathbb{L}_{I_o}$. Using $\varkappa \ge \frac{2\nu+2\alpha+3}{2\alpha}$, the fact that $\theta^T(\mathrm{P}_{\mathbb{L}_I} - \mathrm{P}_{I_0})\theta = \|\mathrm{P}_{\mathbb{L}_I \cap \mathbb{L}_{I_o}^\perp}\theta\|^2 \le \sigma^2\rho(s(I))$ and Lemma 5.1 (in case $\mathbb{L}_{I_0} \subseteq \mathbb{L}_I$) with $h = \alpha$, we obtain for each $I \in \mathcal{G}_0 = \{I \in \mathcal{I} : \rho(s(I)) \ge M_0'\rho(s(I_0)) + M\}$ with $M_0' = 2(\varkappa\alpha + \frac{\alpha}{2})$,

$$\begin{aligned} \mathbb{E}_\theta \hat{p}_I &\le \big(\tfrac{\lambda_I}{\lambda_{I_0}}\big)^\alpha \exp\big\{\alpha\sigma^{-2}\theta^T(\mathrm{P}_{\mathbb{L}_I} - \mathrm{P}_{\mathbb{L}_{I_0}})\theta + \tfrac{\alpha}{2}\rho(s(I_0)) + \rho(s(I))\big\} \\ &\le \exp\big\{-(\varkappa\alpha - \alpha - 1)\rho(s(I)) + (\alpha\varkappa + \tfrac{\alpha}{2})\rho(s(I_0))\big\} \\ &\le \exp\big\{-(\nu + \tfrac{1}{2})\rho(s(I)) + (\alpha\varkappa + \tfrac{\alpha}{2})\rho(s(I_0))\big\} \\ &\le \exp\big\{-\nu\rho(s(I)) - (\tfrac{M_0'}{2} - \varkappa\alpha - \tfrac{\alpha}{2})\rho(s(I_0)) - \tfrac{M}{2}\big\} = e^{-\nu\rho(s(I)) - M/2}. \end{aligned}$$

Combining the last display with Condition (A2) completes the proof:

$$\begin{aligned} \mathbb{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : \rho(s(I)) \ge M_0'\rho(s(I_o)) + M \big| Y\big) &\le \mathbb{E}_\theta \hat{\pi}(I \in \mathcal{G}_0|Y) \\ &= \sum_{I \in \mathcal{G}_0} \mathbb{E}_\theta \hat{p}_I \le C_\nu e^{-M/2}. \end{aligned}$$

□

Proof. of Theorem 5.3 We first establish the coverage property. The constants M_1, H_1 and m_1 are defined in Theorem 5.1. Take $M_2 = \frac{M_1}{\delta}$, where $\delta \in (0,1)$ is from (5.21). From (5.15), it follows that $r^2(\theta) \le r^2(I_*,\theta) = (b(\theta)+1)\sigma^2\rho(s(I_*)) + b(\theta)\sigma^2 \le (b(\theta)+1)\sigma^2(\rho(s(I_*)) + 1)$, where $b(\theta)$ is given by (5.26). Combining this with the claim (5.18) from Theorem 5.1, the claim (ii) from Theorem 5.2 and the definition (5.23) of $\hat{r}$ yields the coverage property:

$$\begin{aligned} &\mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, [(b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2]^{1/2})\big) \\ &\le \mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 > (b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2, \hat{r}^2 \ge \delta\sigma^2\rho(s(I_*)) + \sigma^2 - \tfrac{M\sigma^2}{M_2}\big) \\ &\quad + \mathbb{P}_\theta\big(\hat{r}^2 < \delta\sigma^2\rho(s(I_*)) + \sigma^2 - \tfrac{M\sigma^2}{M_2}\big) \\ &\le \mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 > M_1 r^2(\theta) + M\sigma^2\big) + \mathbb{P}_\theta\big(\rho(s(\hat{I})) < \delta\rho(s(I_*)) - \tfrac{M}{M_2}\big) \\ &\le H_1 e^{-m_1 M} + C_\nu e^{-m_1'' M} \le H_2 e^{-m_2 M}, \end{aligned}$$

where $m_1'' = m_1'/M_2$, $H_2 = H_1 + C_v$, $m_2 = m_1 \wedge m_1''$; m_1' is defined in Theorem 5.2. Since $b(\theta) \le t$ for all $\theta \in \Theta_{\mathrm{eb}}(t)$, the coverage relation follows.

Let us show the size property. For $M \ge 0$, introduce the set $\mathcal{G}(M) = \mathcal{G}(M,\theta) = \{I \in \mathcal{I} : \sigma^2\rho(s(I)) \ge c_3 r^2(\theta) + M\sigma^2\}$, where c_3 is defined in Lemma 5.2. Then for all $I \in \mathcal{G}(M)$,

$$r^2(I,\theta) - c_3 r^2(\theta) \ge \sigma^2\rho(s(I)) - c_3 r^2(\theta) \ge M\sigma^2.$$

Remind the notation (in this theorem) $\hat{p}_I = \check{\pi}(I|Y) = 1\{I = \hat{I}\}$ defined by (5.14). From Lemma 5.2 and the last relation, it follows that for all $I \in \mathcal{G}(M)$

$$\mathbb{E}_\theta \hat{p}_I \le \exp\bigl\{-c_1\rho(s(I)) - c_2\sigma^{-2}\bigl[r^2(I,\theta) - c_3 r^2(\theta)\bigr]\bigr\} \le e^{-c_1\rho(s(I))-c_2 M}.$$

The last display implies that, for any $\theta \in \Theta$,

$$\begin{aligned}
&\mathbb{P}_\theta(\hat{r}^2 \ge c_3 r^2(\theta) + (M+1)\sigma^2) = \mathbb{P}_\theta(\sigma^2\rho(s(\hat{I})) \ge c_3 r^2(\theta) + M\sigma^2) \\
&\le \sum_{I \in \mathcal{G}(M)} \mathbb{E}_\theta \hat{p}_I \le e^{-c_2 M} \sum_{I \in \mathcal{I}} e^{-c_1\rho(s(I))} \le H_3 e^{-c_2 M},
\end{aligned}$$

because $\sum_{I\in\mathcal{I}} e^{-c_1\rho(s(I))} \le C_v$ in view of Condition (A2) as $c_1 > 2v$. The size relation follows with $M_3 = c_3$, $H_3 = C_v$ and $m_3 = c_2$.

If, instead of Condition (A3), stronger Condition (A3') is fulfilled, then the stronger version of the size relation follows immediately from property (iii) of Theorem 5.2: $\mathbb{P}_\theta\bigl(\hat{r}^2 \ge M_0'\sigma^2\rho(s(I_o)) + (M+1)\sigma^2\bigr) \le C_v e^{-M/2}$, where the constants M_0' and C_v are defined in Theorem 5.2.

□

Proof. of Theorem 5.4 Since $Y' = \mathrm{P}_{I^*}\theta + \xi'$, we rewrite (5.27) as

$$\begin{aligned}
\tilde{R}_M^2 &= \bigl(\|Y' - \hat{\theta}\|^2 - \sigma^2 V(Y') + 2\sigma^2 G_M\sqrt{N}\bigr)_+ \\
&= \bigl(\|\theta - \hat{\theta}\|^2 + \sigma^2\bigl(\|\xi'\|^2 - V(Y')\bigr) + 2\sigma\langle\xi', (\theta - \hat{\theta})\rangle + 2\sigma^2 G_M\sqrt{N}\bigr)_+. \qquad (5.43)
\end{aligned}$$

Introduce the events $D_M = D_M(\theta) = \bigl\{\|\hat{\theta} - \theta\|^2 \ge M_1 r^2(\theta) + M\sigma^2\bigr\}$ and $E_M = E_M(\theta) = \bigl\{2|\langle\xi', (\theta - \hat{\theta})\rangle| \ge \sqrt{M(M_1 r^2(\theta) + M\sigma^2)}\bigr\}$. According to Condition (A4), $\hat{\theta}$ and $\hat{I}$ are based on Y and independent of ξ'. Using this fact, the first relation from (A4) and Theorem 5.1, we obtain that

$$\begin{aligned}
&\mathbb{P}_\theta(E_M) = \mathbb{E}_\theta \mathbb{P}_\theta(E_M \cap D_M^c | Y) + \mathbb{P}_\theta(E_M \cap D_M) \\
&\le \mathbb{E}_\theta\Bigl[\psi_1\bigl(\tfrac{M(M_1 r^2(\theta) + M\sigma^2)}{4\|\hat{\theta} - \theta\|^2}\bigr) 1_{D_M^c}\Bigr] + \mathbb{P}_\theta(D_M) \le \psi_1(M/4) + H_1 e^{-m_1 M}. \qquad (5.44)
\end{aligned}$$

Since, by (5.16), $r^2(\theta) \le \sigma^2 N$, the event E_M^c implies that $2\sigma\langle\xi', (\theta - \hat{\theta})\rangle > -\sigma\sqrt{M(M_1\sigma^2 N + M\sigma^2)} \ge -\sigma^2 G_M\sqrt{N}$. Combining this with (5.43), (5.44) and the second relation from (A4) yields the coverage property:

$$\begin{aligned}
&\mathbb{P}_\theta\bigl(\theta \notin B(\hat{\theta}, \tilde{R}_M)\bigr) = \mathbb{P}_\theta\bigl(\theta \notin B(\hat{\theta}, \tilde{R}_M), E_M^c\bigr) + \mathbb{P}_\theta\bigl(\theta \notin B(\hat{\theta}, \tilde{R}_M), E_M\bigr) \\
&\quad \le \mathbb{P}_\theta\bigl(\|\theta - \hat{\theta}\|^2 \ge \tilde{R}_M^2, E_M^c\bigr) + \mathbb{P}_\theta(E_M) \\
&\quad \le \mathbb{P}_\theta\bigl(0 \ge \sigma^2(\|\xi'\|^2 - V(Y')) + \sigma^2 G_M\sqrt{N}\bigr) + \mathbb{P}_\theta(E_M) \\
&\quad \le \mathbb{P}_\theta\bigl(\|\xi'\|^2 - V(Y') \le -M\sqrt{N}\bigr) + \psi_1(M/4) + H_1 e^{-m_1 M} \\
&\quad \le \psi_2(M) + \psi_1(M/4) + H_1 e^{-m_1 M}.
\end{aligned}$$

5

Let us show the size property. By (5.44), $\mathbb{P}_\theta\big(2\sigma\langle\xi',(\theta-\hat\theta)\rangle \ge \sigma^2 G_M\sqrt{N}\big) \le \mathbb{P}_\theta\big(2\langle\xi',(\theta-\hat\theta)\rangle > \sqrt{M(M_1 r^2(\theta)+M\sigma^2)}\big) \le \mathbb{P}_\theta(E_M) \le \psi_1(M/4)+H_1 e^{-m_1 M}$. This, Theorem 5.1 and (5.43) imply

$$\begin{aligned}\mathbb{P}_\theta\big(\tilde R_M^2 \ge g_M(\theta,N)\big) &\le \mathbb{P}_\theta\big(\|\theta-\hat\theta\|^2 \ge M_1 r^2(\theta)+M\sigma^2\big)\\ &\quad+\mathbb{P}_\theta\big(\sigma^2\big(\|\xi'\|^2-V(Y')\big)\ge\sigma^2 G_M\sqrt{N}\big)+\mathbb{P}_\theta\big(2\sigma\langle\xi',(\theta-\hat\theta)\rangle\ge\sigma^2 G_M\sqrt{N}\big)\\ &\le H_1 e^{-m_1 M}+\psi_2(M)+\psi_1(M/4)+H_1 e^{-m_1 M}.\end{aligned}$$

□

5.5. APPLICATIONS

In this section we introduce a number of particular models and structures which fall into the studied general framework and for which local and adaptive (global) minimax results can be derived as consequences of our local results for the general framework.

For all the considered models, we specify the family of structures $\mathcal{I}$, the structural slicing mapping $s:\mathcal{I}\mapsto\mathcal{S}$, and the majorant $\rho(s(I))$. We will further verify Conditions (A1), (A2), (A3) and (A4), whenever appropriate. In view of Remarks 1.6 and 5.10, Conditions (A1) and (A4) hold with $d_{s(I)}=\dim(\mathbb{L}_I)$ in all models with $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. Therefore, we will not verify Conditions (A1) and (A4) for those models.

We keep the same notation for all the quantities involved as for the general framework, with the understanding that these are specialized for the particular models and structures, and some constants must be adjusted. First we summarize the results of Theorems 5.1, 5.2, 5.3 and 5.4 by the following corollary.

Corollary 5.1. *Let Conditions* (A1) *and* (A2) *be fulfilled. Then for any* $M\ge 0$

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta\hat\pi\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta)+M\sigma^2|Y\big) \le H_0 e^{-m_0 M}, \tag{i}$$

$$\sup_{\theta\in\Theta}\mathbb{P}_\theta\big(\|\hat\theta-\theta\|^2 \ge M_1 r^2(\theta)+M\sigma^2\big) \le H_1 e^{-m_1 M}, \tag{ii}$$

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta\hat\pi\big(I\in\mathcal{I}: r^2(I,\theta)\ge c_3 r^2(\theta)+M\sigma^2|Y\big) \le C_\nu e^{-c_2 M}, \tag{iii}$$

$$\sup_{\theta\in\Theta}\mathbb{P}_\theta\big(\hat r^2 \ge M_3 r^2(\theta)+(M+1)\sigma^2\big) \le H_3 e^{-m_3 M}. \tag{iv}$$

If in addition Condition (A3) *is fulfilled, then for any* $M,t\ge 0$

$$\sup_{\theta\in\Theta_{\mathrm{eb}}(t)}\mathbb{P}_\theta\big(\theta\notin B(\hat\theta,\hat R_M)\big) \le H_2 e^{-m_2 M}. \tag{v}$$

If in addition Condition (A4) *is fulfilled, then for any* $M\ge 0$,

$$\sup_{\theta\in\Theta}\mathbb{P}_\theta\big(\theta\notin B(\hat\theta,\tilde R_M)\big) \le \psi_1(M/4)+\psi_2(M)+H_1 e^{-m_1 M}, \tag{vi}$$

$$\sup_{\theta\in\Theta}\mathbb{P}_\theta\big(\tilde R_M^2 \ge g_M(\theta,N)\big) \le \psi_1(M/4)+\psi_2(M)+2H_1 e^{-m_1 M}. \tag{vii}$$

All the constants and quantities are defined in Theorems 5.1, 5.2, 5.3 and 5.4.

Remark 5.11. *The properties (ii) and (iii) of Theorem 5.2 can also be included in Corollary 5.1, but we omit them for the sake of brevity. These are auxiliary results used only for proving the size relation of Theorem 5.3, although one can see these claims as inference on the complexity of the oracle structure.*

If additionally Condition (A3') is assumed for the property (iv), then the stronger uniform version of (iv) holds: $\sup_{\theta\in\Theta}\mathbb{P}_\theta\big(\hat{r}^2 \ge M_0'\rho(s(I_o))\sigma^2+(M+1)\sigma^2\big) \le C_v e^{-M/2}$. *Claim (v) of Corollary 5.1 can be formulated for the local version of the coverage relation of Theorem 5.3 in terms of* $b(\theta)$ *(given by (5.26)) if needed.*

Let $\{\Theta_\beta, \beta\in\mathcal{B}\}$ be a scale of classes, where $\beta\in\mathcal{B}$ is a structural parameter, for instance, β could measure the amount of *smoothness* or *sparsity* of $\theta\in\Theta_\beta$. The local results of Corollary 5.1 imply corresponding global minimax adaptive results over *all* scales $\{\Theta_\beta, \beta\in\mathcal{B}\}$ at once, whose minimax rate

$$r^2(\Theta_\beta) \triangleq \inf_{\tilde\theta}\sup_{\theta\in\Theta_\beta}\mathbb{E}_\theta\|\tilde\theta-\theta\|^2$$

is bounded from below by a multiple of the local rate, namely

$$r^2(\Theta_\beta) \ge c r^2(\theta) \quad \text{for all } \theta\in\Theta_\beta,\ \beta\in\mathcal{B}. \tag{5.45}$$

Remark 5.12. *Typically, (5.45) is established by comparing the oracle rate with the rate for some appropriately chosen structure* $I^* = I^*(\theta)$. *The reasoning goes usually as follows: first show that* $\sup_{\theta\in\Theta_\beta}\|\theta-\mathrm{P}_{I^*}\theta\|^2 \lesssim r^2(\Theta_\beta)$ *and* $\sigma^2\rho(s(I^*)) \lesssim r^2(\Theta_\beta)$, *then argue* $r^2(\theta) \le r^2(I^*,\theta) = \|\theta-\mathrm{P}_{I^*}\theta\|^2+\sigma^2\rho(s(I^*)) \lesssim r^2(\Theta_\beta)$ *uniformly in* $\theta\in\Theta_\beta$. *Often* I^* *is the so called "true structure", i.e.,* $\theta\in\mathbb{L}_{I^*}$, *then* $r^2(I^*,\theta) = \sigma^2\rho(s(I^*)) \lesssim r^2(\Theta_\beta)$.

If (5.45) holds, we say that the oracle rate $r^2(\theta)$ *covers* the scale $\{\Theta_\beta, \beta\in\mathcal{B}\}$. For example, under (5.45), the adaptive (with respect to the structural parameter $\beta\in\mathcal{B}$) minimax estimation result follows immediately from Theorem 5.1: $\sup_{\theta\in\Theta_\beta}\mathbb{P}_\theta\big(\|\hat\theta-\theta\|^2 \ge \frac{M_1}{c}r^2(\Theta_\beta)+M\sigma^2\big) \le H_1 e^{-m_1 M}$. Moreover, Theorems 5.1 and 5.3 imply the adaptive minimax versions of the posterior contraction result, the estimation result and the size relation in the uncertainty quantification problem, which are summarized by the following corollary.

Corollary 5.2. *Let (5.45) and Conditions (A1) and (A2) be fulfilled. Then for any* $M\ge 0$,

$$\sup_{\theta\in\Theta_\beta}\mathbb{E}_\theta\hat\pi\big(\|\vartheta-\theta\|^2 \ge M_0 c^{-1}r^2(\Theta_\beta)+M\sigma^2\,|\,Y\big) \le H_0 e^{-m_0 M},$$

$$\sup_{\theta\in\Theta_\beta}\mathbb{P}_\theta\big(\|\hat\theta-\theta\|^2 \ge M_1 c^{-1}r^2(\Theta_\beta)+M\sigma^2\big) \le H_1 e^{-m_1 M},$$

$$\sup_{\theta\in\Theta_\beta}\mathbb{P}_\theta\big(\hat r^2 \ge M_3 c^{-1}r^2(\Theta_\beta)+M\sigma^2\big) \le H_3 e^{-m_3 M}.$$

In case the radius of confidence ball is of the order $r(\theta)+\sigma N^{1/4}$, we assume that the conditions of Theorem 5.4 instead of Theorem 5.3 are fulfilled and the third claim of Corollary 5.2 is replaced as follows:

$$\sup_{\theta\in\Theta_\beta}\mathbb{P}_\theta\big(\tilde R_M^2 \ge g_M'(\theta,N)\big) \le \psi_1(M/4)+\psi_2(M)+2H_1 e^{-m_1 M},$$

5

where $g'_M(\theta, N) = M_1 c^{-1} r^2(\Theta_\beta) + M\sigma^2 + 4\sigma^2 G_M \sqrt{N}$.

We do not specialize Theorem 5.2 and the coverage relation of Theorems 5.3 and 5.4 for the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$, because it does not make much sense to specialize these claims for any scale (the minimax risk is not present in these claims). Theorem 5.2 holds uniformly in $\theta \in \Theta$, hence uniformly over any Θ_β. The coverage relation in Theorem 5.3 holds uniformly over the EBR class Θ_{eb}, so it will certainly hold uniformly over the intersection $\Theta_{\text{eb}} \cap \Theta_\beta$. Similarly, the coverage relation in Theorem 5.4 will certainly hold uniformly over Θ_β.

Below we verify the required conditions to obtain Corollaries 5.1 and 5.2 for concrete models and structures. For brevity sake, for some examples and some claims of Corollaries 5.1 and 5.2, we will not present all the computations for verifying the required conditions, since these computations can be done similarly to the previously considered cases.

5.5.1. SIGNAL+NOISE MODEL WITH SMOOTHNESS STRUCTURE

Assume that the data $Y = (Y_i)_{i\in\mathbb{N}}$ come from the model

$$Y_i = \theta_i + \tfrac{1}{\sqrt{n}}\xi_i, \;\; i \in \mathbb{N},$$

where $\theta = (\theta_i)_{i\in\mathbb{N}} \in \Theta = \ell_2$ is an unknown parameter and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. A local approach for this model, delivering also the adaptive minimax results for many smoothness structures simultaneously, is considered by [5], [42] for posterior contraction rates and by [9] for uncertainty quantification problem.

Admittedly, this is an infinite dimensional model as compared with the default high-dimensional general framework (1.15), but in this case all the results go through with one minor adjustment: all the sums over $I \in \mathcal{I}$ become countable infinite instead of finite. Alternatively, we could consider a finite dimensional model approximating the original infinite dimensional model with arbitrary accuracy.

In this case, the smoothness structure is modeled by the linear spaces

$$\mathbb{L}_I = \{x \in \ell_2 : x_i = 0 \text{ for all } i \geq I+1\}, \quad I \in \mathcal{I} = \mathbb{N}_0. \tag{5.46}$$

We have $\|\theta - \mathrm{P}_I\theta\|^2 = \sum_{i=I+1}^\infty \theta_i^2$, $d_I = \dim(\mathbb{L}_I) = I$, the structural slicing mapping is taken to be $s(I) = I$, so that $\mathcal{S} = \mathcal{I} = \mathbb{N}_0$ and $\mathcal{I}_{s(I)} = \{I\}$. Hence $\log|\mathcal{I}_s| = 0$ for all $s \in \mathcal{S}$, and we thus take the majorant $\rho(s(I)) = d_{s(I)} + \log|\mathcal{I}_{s(I)}| = d_I = I$. The oracle rate is

$$r^2(\theta) = \min_{I\in\mathbb{N}_0}\big(\textstyle\sum_{i\geq I+1}\theta_i^2 + \frac{I}{n}\big) = \sum_{i\geq I_o+1}\theta_i^2 + \frac{I_o}{n}.$$

Recall that, in view of Remarks 1.6 and 5.10, Conditions (A1) and (A4) hold with $d_{s(I)} = \dim(\mathbb{L}_I)$. Condition (A2) is fulfilled since, in view of Remark 1.8, $\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} = \sum_{s\in\mathcal{S}} e^{-\nu s} = \frac{e^\nu}{e^\nu - 1} = C_\nu$ for any $\nu > 0$. Finally, Condition (A3) is also fulfilled. Indeed, for any $I_0, I_1 \in \mathcal{I}$ define $I'(I_0, I_1) = I_0 \vee I_1$, then $(\mathbb{L}_{I_0} \cup \mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'}$ and $\rho(s(I')) = I_0 \vee I_1 \leq I_0 + I_1 = \rho(s(I_0)) + \rho(s(I_1))$.

As consequence of our general results, we obtain the local results of Corollary 5.1 for this case with the local rate $r^2(\theta)$ defined above. In turn, by virtue of Corollary 5.2 the local results will imply global minimax adaptive results at once over all scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$ (i.e., for which (5.45) holds). Below we present a couple of examples of scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$.

Sobolev ellipsoids. For $\beta, Q > 0$, introduce the Sobolev ellipsoids

$$\Theta_\beta = \Theta_\beta(Q) = \{\theta \in \ell_2 : \textstyle\sum_{i\in\mathbb{N}} i^{2\beta}\theta_i^2 \le Q\}. \tag{5.47}$$

It is well known that the corresponding minimax rate is $r^2(\Theta_\beta) = n^{-2\beta/(2\beta+1)}$; see, for example, [12] or [71]. The adaptive minimax results for Sobolev ellipsoids were considered by [5], [82] (see further references therein) for posterior contraction rates, and by [9], [77], [83] (see further references therein) for constructing optimal confidence balls. By taking $I_0 = \lfloor n^{1/(2\beta+1)} \rfloor$, we obtain (5.45):

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} r^2(\theta) &= \sup_{\theta\in\Theta_\beta(Q)} \Big(\sum_{i=I_o+1}^{\infty} \theta_i^2 + \tfrac{I_o}{n} \Big) \le \sup_{\theta\in\Theta_\beta(Q)} \sum_{i=I_0+1}^{\infty} \tfrac{i^{2\beta}\theta_i^2}{I_0^{2\beta}} + \tfrac{I_0}{n} \\ &\le \tfrac{I_0}{n} + \tfrac{Q}{I_0^{2\beta}} \lesssim n^{-2\beta/(2\beta+1)} \asymp r^2(\Theta_\beta).\end{aligned}$$

Corollary 5.2 follows for this case with the minimax rate $r^2(\Theta_\beta)$ defined above.

Sobolev hyperrectangles. Consider the so called hyperrectangles in ℓ_2:

$$\Theta_\beta = \Theta_\beta(Q) = \{\theta \in \ell_2 : |\theta_i| \le \sqrt{Q} i^{-\beta}\}, \ \beta > 1/2.$$

It is not difficult to show that the corresponding minimax rate is $r^2(\Theta_\beta) = n^{-(2\beta-1)/(2\beta)}$. The adaptive minimax results for Sobolev hyperrectangles were considered by [5], [9] for posterior contraction rates, and by [9], [77], [83] (see further references therein) for constructing optimal confidence balls. By taking $I_0 = \lfloor n^{1/2\beta} \rfloor$, we obtain (5.45):

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} r^2(\theta) &= \sup_{\theta\in\Theta_\beta(Q)} \sum_{i=I_o+1}^{\infty} \theta_i^2 + \tfrac{I_o}{n} \le \sup_{\theta\in\Theta_\beta(Q)} \sum_{i=I_0+1}^{\infty} \tfrac{Q}{i^{2\beta}} + \tfrac{I_0}{n} \\ &\le \tfrac{I_0}{n} + \tfrac{Q}{(2\beta-1)I_0^{(2\beta-1)}} \lesssim n^{-(2\beta-1)/(2\beta)} \asymp r^2(\Theta_\beta).\end{aligned}$$

Corollary 5.2 follows for this case with the minimax rate $r^2(\Theta_\beta)$ defined above.

Analytic and tail classes. Similarly, we can derive the adaptive minimax results for two more scales of *exponential ellipsoids* (or *analytic classes*) and *tail classes*. Exponential ellipsoids are defined as follows:

$$\Theta_\beta = \Theta_\beta(Q) = \{\theta \in \ell_2 : \textstyle\sum_{k\in\mathbb{N}} e^{2\beta k}\theta_k^2 \le Q\}, \quad \beta > 0.$$

For the analytic scale, the relation (5.45) is $\sup_{\theta\in\Theta_\beta} r^2(\theta) \lesssim r^2(\Theta_\beta) \asymp \frac{\log n}{n}$.

The tail classes are

$$\Theta_\beta = \Theta_\beta(Q) = \{\theta \in \ell_2 : \textstyle\sum_{k=m}^{\infty} \theta_k^2 \le Q m^{-\beta}, m \in \mathbb{N}\}, \quad \beta > 0.$$

In this case, the relation (5.45) is $\sup_{\theta\in\Theta_\beta} r^2(\theta) \lesssim r^2(\Theta_\beta) \asymp n^{-\beta/(\beta+1)}$.

Corollary 5.2 follows for the both scales with the corresponding minimax rates $r^2(\Theta_\beta)$ defined above.

5.5.2. SIGNAL+NOISE MODEL UNDER WAVELET BASIS

We adopt the notation and conventions from [49]. Consider the observations

$$Y_{jk} = \theta_{jk} + \tfrac{1}{\sqrt{n}}\xi_{jk}, \quad \xi_{jk} \overset{\text{ind}}{\sim} \mathrm{N}(0,1), \quad (jk) \in \mathcal{K} = \{(jk) : j \in \mathbb{N}_0,\ k \in [2^j]\}.$$

This model is obtained as the result of the orthogonal wavelet transform of an additive regression function observed in Gaussian noise with $\sigma^2 = n^{-1}$, or just as a sequence version (with respect to some wavelet basis) of the continuous *white noise model*. We could also consider a high dimensional "projected" (see (9.57) in [49]) variant, where $j \in [J_o]$ with $2^{J_0+1} = n$. For details and many interesting connections and relations to the function estimation theory we refer to the very comprehensive and insightful account [49] on this topic.

The smoothness structure of $\theta = (\theta_{jk}, (jk) \in \mathcal{K})$ is modeled by the linear spaces

$$\mathbb{L}_I = \big\{(x_{jk}, (jk) \in \mathcal{K}) : x_{jk} = 0 \ \forall\, j \in [j_0]_0, k \in I_j^c \text{ and } \forall\, j > j_0,\ k \in [2^j]\big\},$$

where $I = (j_0, I_0, \ldots, I_{j_0})$ with $I_j \subseteq [2^j]$, $j \in [j_0]_0$. The structural slicing mapping is $s(I) = (j_0, |I_0|, \ldots, |I_{j_0}|)$ and $d_I = \dim(\mathbb{L}_I) = \sum_{m=0}^{j_0} |I_m|$. Compute $|\mathcal{I}_{s(I)}| = \prod_{k=0}^{j_0} \binom{2^k}{|I_k|}$, hence $\log|\mathcal{I}_{s(I)}| = \sum_{k=0}^{j_0} \log\binom{2^k}{|I_k|} \le \sum_{k=0}^{j_0} |I_k| \log(\frac{e2^k}{|I_k|})$. Since $d_{s(I)} + \log|\mathcal{I}_{s(I)}| = d_I + \log|\mathcal{I}_{s(I)}| \le 2\sum_{k=0}^{j_0} |I_k| \log(\frac{e2^k}{|I_k|})$, we take the majorant $\rho(s(I)) = 2\sum_{k=0}^{j_0} |I_k| \log(\frac{e2^k}{|I_k|})$.

Conditions (A1) and (A4) hold with $d_{s(I)} = \dim(\mathbb{L}_I)$ in view of Remarks 1.6 and 5.10. Condition (A2) is also fulfilled, since, according to Remark 1.8, for any $\nu > 2$

$$\begin{aligned}\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} &\le \sum_{s\in\mathcal{S}} e^{-(\nu-1)\rho(s(I))} \le \sum_{j_0=0}^{\infty} \sum_{k_0=1}^{2^0} \cdots \sum_{k_m=1}^{2^{j_0}} e^{-(\nu-1)(k_0+\ldots+k_m)} \\ &\le \sum_{j_0=0}^{\infty} \big(\tfrac{1}{e^{\nu-1}-1}\big)^{j_0+1} \le \tfrac{1}{e^{\nu-1}-2} = C_\nu.\end{aligned}$$

Finally, for any $I^0, I^1 \in \mathcal{I}$ define $j_0'' = \min\{j_0^0, j_0^1\}$, $j_0' = \max\{j_0^0, j_0^1\}$ and $I'(I^0, I^1) \in \mathcal{I}$ such that

$$I'(I^0, I^1) = (I_0^0 \cup I_0^1, I_1^0 \cup I_1^1, \ldots, I_{j_0''}^0 \cup I_{j_0''}^1, I_{j_0''+1}^{1\{j_0'=j_0^1\}}, \ldots, I_{j_0'}^{1\{j_0'=j_0^1\}}).$$

Then $(\mathbb{L}_{I^0} \cup \mathbb{L}_{I^1}) \subseteq \mathbb{L}_{I'}$ and

$$\sum_{m=0}^{j_0'} |I'_m| \log\big(\tfrac{e2^m}{|I'_m|}\big) \le \sum_{m=0}^{j_0^0} |I_m^0| \log\big(\tfrac{e2^m}{|I_m^0|}\big) + \sum_{m=0}^{j_0^1} |I_m^1| \log\big(\tfrac{e2^m}{|I_m^1|}\big),$$

which entails Condition (A3).

As consequence of our general results, we obtain Corollary 5.1 for this case with the local rate $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|\theta - \mathrm{P}_I\theta\|^2 + \frac{1}{n}\rho(s(I))\}$. Below we present the example of Besov scale, for which the global minimax adaptive results follow from the local results. We should mention that there are of course more scales covered by the oracle rate $r^2(\theta)$, the reader is invited to make computations for other interesting scales. Besides, the results can be extended to non-normal, not independent ξ_{jk}'s, but only satisfying Condition (A1).

Besov scale. Assume that the true signal θ belongs to a Besov ball

$$\Theta^{\beta}_{p,q}(Q) = \Big\{\theta : \sum_{j=0}^{\infty} 2^{ajq}\big(\sum_{k=1}^{2^j} \theta_{jk}^p\big)^{q/p} \le Q^q\Big\}, \quad a = \beta + \frac{1}{2} - \frac{1}{p},$$

for some $p, q, Q > 0$ and $\beta \ge 1/p$. The minimax rate over $\Theta^{\alpha}_{p,q}(Q)$ is known to be $r^2(\Theta^{\beta}_{p,q}(Q)) \asymp n^{-\frac{2\beta}{2\beta+1}}$. The adaptive minimax results for the scale of the class $\Theta^{\beta}_{p,q}(Q)$ were considered by [87], [76], [48], [42] and many others for posterior contraction rates, and [26] for constructing optimal confidence balls.

Let $j_* = \lfloor \log_2 n \rfloor$. Define $\mathcal{I}_* = \{I \in \mathcal{I} : j_0(I) = j_*\}$ and note that $\mathcal{I}_* \subset \mathcal{I}$. Hence, for any $\theta \in \Theta^{\beta}_{p,q}(Q)$,

$$\begin{aligned} r^2(\theta) &\le \min_{I \in \mathcal{I}_*} \big\{\|\theta - \mathrm{P}_I\theta\|^2 + \tfrac{1}{n}\rho(s(I))\big\} \\ &\le \sum_{j=0}^{j_*} \sum_{k \in I^c_{oj}} \theta^2_{jk} + \sum_{j=j_*+1}^{\infty} \sum_{k=1}^{2^j} \theta^2_{jk} + \sum_{j=0}^{j_*} \tfrac{|I_{oj}|}{n} \log\big(\tfrac{e2^j}{|I_{oj}|}\big) \\ &\le \sum_{j=0}^{j_*} \min_{0 \le k \le 2^j} \Big(\sum_{l>k} \theta^2_{j(l)} + C_1 \frac{k}{n} \log(e2^j/k)\Big) + \sum_{j=j_*+1}^{\infty} \sum_{k=1}^{2^j} \theta^2_{jk} \\ &\le C_2 n^{-\frac{2\beta}{2\beta+1}} + C_3 n^{-1} \lesssim n^{-\frac{2\beta}{2\beta+1}} \asymp r^2(\Theta^{\beta}_{p,q}(Q)), \end{aligned}$$

where $\theta^2_{j(l)}$ denotes the l-th largest value among $\{\theta^2_{jk}, j \in [2^k]\}$. The third inequality of the last display follows from Theorem 12.1 in [49] under the assumption $\beta \ge 1/p$. We thus established the relation (5.45) for the Besov scale, and Corollary 5.2 follows with the minimax rate $r^2(\Theta^{\beta}_{p,q}(Q))$ defined above.

5.5.3. SIGNAL+NOISE MODEL WITH (MULTI-LEVEL) SPARSITY STRUCTURE

Assume that the data $Y = (Y_i)_{i \in [n]}$ come from the model

$$Y_i = \theta_i + \sigma\xi_i, \ \ i \in [n], \tag{5.48}$$

where $\theta = (\theta_i)_{i \in [n]} \in \Theta = \mathbb{R}^n$ is an unknown parameter and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. The high-dimensional vector θ is assumed to be *sparse*. A local approach for this model, delivering also the adaptive minimax results for various sparsity structures simultaneously, is considered in [13], [42] for posterior contraction rates and by [13] for uncertainty quantification problem.

The classical sparsity structure is modeled by the linear spaces

$$\mathbb{L}_I = \big\{x \in \mathbb{R}^n : x_i = 0, i \in I^c\big\}, \quad I \in \mathcal{I} = \{J : J \subseteq [n]\}.$$

In this case, $d_I = \dim(\mathbb{L}_I) = |I|$, $\|\theta - \mathrm{P}_I\theta\|^2 = \sum_{i \in I^c} \theta_i^2$, the structural slicing mapping is defined to be $s(I) = |I| \in \mathcal{S} \triangleq [n]_0$. Compute $|\mathcal{I}_{s(I)}| = \binom{n}{|I|}$, hence $\log|\mathcal{I}_{s(I)}| = \log\binom{n}{|I|} \le |I|\log(\frac{en}{|I|})$. Since $d_{s(I)} + \log|\mathcal{I}_{s(I)}| = d_I + \log|\mathcal{I}_{s(I)}| \le |I| + |I|\log(\frac{en}{|I|})$, we take the majorant $\rho(s(I)) = 2|I|\log(\frac{en}{|I|})$.

Conditions (A1) and (A4) hold with $d_{s(I)} = \dim(\mathbb{L}_I)$ in view of Remarks 1.6 and 5.10. Condition (A2) is fulfilled, since, according to Remark 1.8, for any $\nu > 1$

$$\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} \le \sum_{s\in\mathcal{S}} e^{-(\nu-1)\rho(s(I))} \le \sum_{s=0}^{n} \left(\tfrac{en}{s}\right)^{-(\nu-1)s} \le \tfrac{1}{1-e^{1-\nu}} = C_\nu.$$

Finally, for any $I_0, I_1 \in \mathcal{I}$ define $I' = I_0 \cup I_1$. Then $(\mathbb{L}_{I_0} \cup \mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'} = \mathbb{L}_{I_0} + \mathbb{L}_{I_1}$ and $|I'|\log\left(\frac{en}{|I'|}\right) \le |I_0|\log\left(\frac{en}{|I_0|}\right) + |I_1|\log\left(\frac{en}{|I_1|}\right)$, which entails Condition (A3).

Remark 5.13. *We can take a slightly better majorant,* $\rho'(s) = \max\{s, \log\binom{n}{s}\}$.

As a consequence of our general results, we obtain Corollary 5.1 with the local rate $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho(s(I))\}$. In view of Remark 5.13, the results hold also with the local rate $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho'(s(I))\}$. As $\rho'(s) \le \rho(s)$ for all $s \in \mathcal{S}$, the local rate with $\rho'(s)$ is smaller than the rate with $\rho(s)$ implying a stronger version of Corollary 5.1. However, the quantity $\rho(s)$ is easier to compute and we will thus use the majorant $\rho(s)$.

There exists a couple of examples of scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ with sparsity structure, for which the global minimax adaptive results follow from the local results, namely, for *nearly black vectors* $\ell_0[p_n]$ and *weak* ℓ_q*-balls* $m_q[p_n]$, which are introduced in Section 2.2.5. In that section, relation (5.45) was established for the scales $\ell_0[p_n]$ and $m_q[p_n]$, so that Corollary 5.2 follows with corresponding minimax rates $r^2(\ell_0[p_n])$ and $r^2(m_q[p_n])$ (defined in Section 2.2.5).

Recall that the results can be extended to non-normal and not necessarily independent ξ_i's, but only satisfying Condition (A1). For example, as demonstrated in [13], ξ_i's originating from a certain AR(1)-model also satisfy Condition (A1).

Multi-level sparsity structure. Consider the same model (5.48), but now with the so called *multi-level sparsity* structure, an extension of the traditional sparsity structure. In the usual one-level sparsity structure we have just one known sparsity level, which is by default zero. The first attempt to study a version of such structure has been undertaken in [14], here we propose a systematic approach to this from the general perspective of the linear spaces for the first time. To the best of our knowledge, this structure has never been systematically studied in the literature.

First we extend the classical sparsity structure by allowing the sparsity level to be an unknown constant, not necessarily zero. The practical implementation of this case is considered in [18]. This extended *unknown level sparsity* structure is described by the linear spaces:

$$\mathbb{L}_I = \left\{x \in \mathbb{R}^n : x_i = x_j, \forall\, i, j \in I^c\right\}, \quad I \in \mathcal{I} = \{J : J \subseteq [n]\}.$$

Then $d_I = \dim(\mathbb{L}_I) = (|I|+1) \wedge n$, $\|\theta - \mathrm{P}_I\theta\|^2 = \sum_{i\in I^c}(\theta_i - \bar\theta_{I^c})^2$ (where $\bar\theta_{I^c} = |I^c|^{-1}\sum_{i\in I^c}\theta_i$), and the structural slicing mapping $s(I) = |I| \in \mathcal{S} \triangleq [n]_0$. Compute $|\mathcal{I}_s| = \binom{n}{s}$, hence $d_{s(I)} + \log|\mathcal{I}_{s(I)}| = d_I + \log\binom{n}{|I|} \le (|I|+1) \wedge n + |I|\log(\frac{en}{|I|})$ and the majorant is $\rho(s(I)) = (|I|+1) \wedge n + |I|\log(\frac{en}{|I|})$.

Next, we extend the one-level sparsity structure to the multi-level sparsity structure (with unknown sparsity levels) by introducing the following linear spaces: for a partition

$I = (I_i, i \in [m]_0)$ of the set $[n]$ into $m+1$ parts,

$$\mathbb{L}_I = \{x \in \mathbb{R}^n : x_j = x_{j'}, \forall\, j, j' \in I_i, i \in [m]\}, \quad I \in \mathcal{I},$$

where $\mathcal{I} = \mathcal{I}_m$ is the family of all partitions of $[n]$ into $m+1$ parts (some possibly empty), and $m = 2, \dots, n-1$. In this case, compute $\|\theta - \mathrm{P}_I\theta\|^2 = \sum_{k=1}^m \sum_{i\in I_k} (\theta_i - \bar\theta_{I_k})^2$ with the group averages $\bar\theta_{I_k} = \frac{1}{|I_k|}\sum_{i\in I_k}\theta_i$, the structural slicing mapping is taken to be $s(I) = (|I_i|, i \in [m]_0) \in \mathcal{S}$, where $\mathcal{S} = \mathcal{S}(n, m+1) = \{(n_i, i \in [m]_0) : n_i \in [n]_0, \sum_{i\in[m]_0} n_i = n\}$ is the family of the so called *weak compositions* of n into $m+1$ parts. It is well known that $|\mathcal{S}| = \binom{n+m}{m}$. Further we have $d_{s(I)} = d_I = \dim(\mathbb{L}_I) = (|I_0| + m) \wedge n$ and $|\mathcal{I}_{s(I)}| = \binom{n}{|I_0|,\dots,|I_m|}$ is the multinomial coefficient.

To ensure Condition (A2), we have to compensate for the number $|\mathcal{S}|$ by adding the term $\log|\mathcal{S}| = \log\binom{n+m}{m}$ in the complexity majorant $\rho(s)$. Hence, we take the majorant $\rho(s(I)) = (|I_0| + m) \wedge n + \log\binom{n}{|I_0|,\dots,|I_m|} + \log\binom{n+m}{m}$.

Conditions (A1) and (A4) hold with $d_{s(I)} = \dim(\mathbb{L}_I)$ in view of Remarks 1.6 and 5.10. Condition (A2) is fulfilled, since for any $\nu \geq 1$

$$\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} = \sum_{s\in\mathcal{S}} \sum_{I\in\mathcal{I}_s} e^{-\nu\rho(s(I))} \leq \sum_{s\in\mathcal{S}} e^{-\nu\log|\mathcal{S}|} \leq 1.$$

Unfortunately, we were unable to establish Condition (A3) for this structure, which is needed for the uncertainty quantification results under the EBR condition. What we can claim are the relations (i)-(iv) and (vi)-(vii) of Corollary 5.1 with the local rate $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho(s(I))\} = \min_{I\in\mathcal{I}}\{\sum_{k=1}^m \sum_{i\in I_k}(\theta_i - \bar\theta_{I_k})^2 + \sigma^2[(|I_0|+m)\wedge n + \log\binom{n}{|I_0|,\dots,|I_m|} + \log\binom{n+m}{m}]\}$. For $m = 1$ we get the classical one-level local sparsity results which also imply the global minimax results over sparsity scales, as is considered in the previous paragraph. For $m \geq 2$, the obtained local results (i)-(iv) and (vi)-(vii) of Corollary 5.1 are new to the best of our knowledge. The most problematic term is $\log\binom{n}{|I_0|,\dots,|I_m|}$, this term of a smaller order than n if $|I_0|$ and any $m-1$ values among $|I_1|, \dots, |I_m|$ (e.g., $|I_0|, |I_1|, \dots, |I_{m-1}|$) are themselves of the smaller order than n.

Remark 5.14. *It is an open problem to establish Condition (A3). This is important in the uncertainty quantification problem, namely, the coverage relation (v) from Corollary 5.1 relies on this condition.*

If we are to verify Condition (A3), for any $I, I' \in \mathcal{I}$ we would define

$$I'' = I''(I, I') = \big(I_0 \cup I'_0, (I_i \cap I'_{i'},\ i, i' \in [m])\big).$$

Clearly, $\mathbb{L}_{I'} \subseteq \mathbb{L}_{I''}$, $\mathbb{L}_I \subseteq \mathbb{L}_{I''} \subseteq \mathbb{L}_I + \mathbb{L}_{I'}$ and $\max\{s(I), s(I')\} \leq s(I'') \leq s(I) + s(I')$, implying $\rho(s(I'')) \leq \rho(s(I)) + \rho(s(I'))$, and seems that Condition (A3) is fulfilled. However, the problem is that the resulting I'' may in general not lie in $\mathcal{I}$ but rather in $\mathcal{I}_{m^2}$. An idea to fix this would be to let the number m of parts in partitions $I \in \mathcal{I}$ free (any integer from 0 to n). But then the problem will emerge in another place: there are too many choices as the family $\mathcal{S}$ of all compositions of n becomes $|\mathcal{S}| = 2^{n-1}$. Then we will have to put the term $\log|\mathcal{S}| \asymp n$ in the complexity majorant $\rho(s(I))$ to meet Condition (A2), which makes the local rate $r^2(\theta) \gtrsim n\sigma^2$ trivially large and therefore uninteresting.

Remark 5.15. *We should mention that the global minimax multi level sparsity results are not going to be useful, at least if we try to extend one-level sparsity scales to multi-level sparsity scales in the usual way. Indeed, even if we assume sparsity in the sense that* $|I_0| \le s$ *for some small* $s = s_n \ll n$*, i.e.,* $\theta \in \Theta_s = \cup_{I\in\mathcal{I}:|I_0|\le s}\mathbb{L}_I$*, the minimax rate over* Θ_s *will presumably be* $r^2(\Theta_s) \asymp \sigma^2 \max_{I\in\mathcal{I}:|I_0|\le s}\rho(s(I)) \gtrsim \sigma^2 \max_{I\in\mathcal{I}:|I_0|\le s}\log\binom{n}{|I_0|,\dots,|I_m|} \gtrsim n\sigma^2$*, and this is not a useful result. This means that the multilevel counterpart* Θ_s *for the traditional one-level sparsity class* $\ell_0[s]$ *is too "massive" in the minimax sense.*

One can propose other scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ *with more structure for which at least minimax consistency would hold, i.e.,* $r^2(\Theta_\beta) \ll \sigma^2 n$*. For example, consider* $\bar\Theta_{s,m} = \cup\{\mathbb{L}_I, I \in \mathcal{I}, i \in [m] : |I_i| \le s, i \in [m]_0\backslash\{i\}\}$*, with* $s = s_n \ll n$ *and* $m = m_n \ll n$*. Then we can show that for any* $\theta \in \bar\Theta_{s_n,m_n}$

$$r^2(\theta) \le \sigma^2\left[(|I_0| + m_n) + \log\binom{n}{|I_0|,\dots,|I_m|} + \log\binom{n+m_n}{m_n}\right] \lesssim \sigma^2(s_n(m_n+1)+m_n)\log n.$$

5.5.4. NOISY FUNCTION ON A LARGE GRAPH WITH SMOOTHNESS STRUCTURE

We adopt the notation and conventions from [51]. Let G be a connected, simple (i.e., no loops, multiple edges or weights), undirected graph with n vertices labelled as $1,\dots,n$. A function f on the (vertices of the) graph is a mapping $f : [n] \mapsto \mathbb{R}$. We will write f both for the function and for the associated vector of function values $(f(1), f(2), ..., f(n))$ in $\mathbb{R}^n$. We assume that $Y_1,\dots,Y_n$ are the observations at the vertices of the graph, satisfying

$$Y_i = f(i) + \tfrac{1}{\sqrt{n}}\xi_i, \quad i \in [n],$$

where $f = (f(i))_{i\in[n]} \in \mathbb{R}^n$ is the vector of values of unknown function on the graph G and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. In this case, $\theta = f \in \Theta \triangleq \mathbb{R}^n$. To the best of our knowledge, there are no local results on estimation, posterior contraction rate and uncertainty quantification problems for this model.

The smoothness structure of function f is described by the linear spaces

$$\mathbb{L}_I = \{x \in \mathbb{R}^n : x_i = 0 \text{ for all } i = I+1\dots n\}, \quad I \in \mathcal{I} = [n]_0.$$

In this case, $\|f - \mathrm{P}_I f\|^2 = \sum_{i=I+1}^n f^2(i)$, the structural slicing mapping $s(I) = I$, so that $\mathcal{S} = \mathcal{I} = [n]_0$ and $\mathcal{I}_s = \mathcal{I}_I = \{I\}$. Hence $\log|\mathcal{I}_s| = 0$. Further, in view of Remark 1.6, Condition (A1) is fulfilled with $\alpha = 0.4$, $d_{s(I)} = d_I = \dim(\mathbb{L}_I) = I$, and we arrive at the majorant $\rho(s(I)) = \rho(I) = d_I = I$. The oracle rate is

$$r^2(f) = \min_{I\in[n]_0}\left(\sum_{i=I+1}^n f^2(i) + \sigma^2 I\right) = \sum_{i=I_o+1}^n f^2(i) + \sigma^2 I_o.$$

Further, Condition (A4) holds in view of Remark 5.10. Condition (A2) is fulfilled since, according to Remark 1.8, for any $\nu > 0$,

$$\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} = \sum_{s\in\mathcal{S}} e^{-\nu s} = \frac{e^\nu}{e^\nu - 1} = C_\nu.$$

Condition (A3) is also fulfilled. Indeed, for any $I_0, I_1 \in \mathcal{I}$ define $I'(I_0, I_1) = I_0 \vee I_1$, then $(\mathbb{L}_{I_0} \cup \mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'}$ and $\rho(s(I')) = I_0 \vee I_1 \le I_0 + I_1 = \rho(s(I_0)) + \rho(s(I_1))$.

As consequence of our general results, we obtain the local results of Corollary 5.1 for this case with the local rate $r^2(f)$ defined above. In turn, by virtue of Corollary 5.2 the local results will imply global minimax adaptive results over all scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ at once, covered by the oracle rate $r^2(f)$ (i.e., for which (5.45) holds). Below we present the Laplacian scale $\{H^\beta, \beta > 0\}$ and show that it is covered by the oracle rate $r^2(f)$.

Minimax results for a Laplacian graph. One common approach to learn functions on graphs is Laplacian regularisation; see, for example, [19] and [51]. The graph Laplacian is defined as $\mathrm{L} = \mathrm{D} - \mathrm{A}$, where A is the adjacency matrix of the graph and D is the diagonal matrix with the degrees of the vertices on the diagonal. When viewed as a linear operator, the Laplacian acts on a function f as

$$\mathrm{L}f(i) = \sum_{j\sim i} \big(f(i) - f(j)\big),$$

where we write $i \sim j$ if vertices i and j are connected by an edge. Denote the Laplacian eigenvalues, ordered by magnitude, by $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$. As in [51], we assume without loss of generality that there exist $i_0 \in \mathbb{N}$, $C_1 > 0$ such that for all n large enough and $r \ge 1$,

$$\lambda_i \ge C_1 \big(\tfrac{i}{n}\big)^{2/r}, \quad i > i_0,$$

and $f \in H^\beta = H^\beta(Q) = \{f : \sum_{i=1}^n (1 + n^{2\beta/r}\lambda_i^\beta) f^2(i) \le Q\}$, with smoothness $\beta > 0$. The minimax estimation rate over the class H^β is $r^2(H^\beta) = \inf_{\tilde f} \sup_{f\in H^\beta} \mathbb{E}_f \|\tilde f - f\|^2 \asymp n^{-\frac{2\beta}{2\beta+r}}$; see [52].

By taking $I_0 = \lfloor n^{r/(2\beta+r)} \rfloor$, we establish (5.45) in this case:

$$\begin{aligned}\sup_{f\in H^\beta} r^2(f) &= \sup_{f\in H^\beta} \sum_{i=I_o+1}^n f^2(i) + \tfrac{I_o}{n} \le \sup_{f\in H^\beta} \sum_{i=I_0+1}^n f^2(i) + \tfrac{I_0}{n} \\ &\le \tfrac{Q}{1+n^{2\beta/r}\lambda_{I_0}^\beta} + \tfrac{I_0}{n} \lesssim n^{-\frac{2\beta}{2\beta+r}} \asymp r^2(H^\beta).\end{aligned}$$

Hence, Corollary 5.2 follows for this case with the minimax rate $r^2(H^\beta)$ defined above. As compared to the Theorems 3.2 and 3.3 in [51], there is no restriction on the range of the smoothness β in Corollary 5.2, and we do not have any extra logarithmic factor in the rate.

5.5.5. DENSITY ESTIMATION WITH SMOOTHNESS STRUCTURE

We observe $X_1, \ldots, X_n \sim f$, where f is a density on $[0,1]$. Let $\{\varphi_i, i \in \mathbb{N}\}$ be an orthonormal basis in $L_2[0,1]$. For simplicity consider a bounded basis $\{\varphi_i, i \in \mathbb{N}\}$, implying $\sup_{x\in[0,1]} |\varphi_i(x)| \le c_\varphi$ for some $c_\varphi > 0$ (e.g., for the trigonometric basis $c_\varphi = \sqrt{2}$). We can expand the density function f in a Fourier series $f(x) = \sum_{i=1}^\infty \theta_i \varphi_i(x)$, $x \in [0,1]$, in the L_2-sense. Due to Parseval's identity, the problem of estimating the density function f in the L_2-sense can be converted into the problem of estimating the parameter $\theta = (\theta_i)_{i\in\mathbb{N}}$ in the ℓ_2-sense:

$$Y_i = \theta_i + \sigma_n \xi_i, \ \ i \in \mathbb{N}, \tag{5.49}$$

where $\theta = (\theta_i)_{i\in\mathbb{N}}$ is an unknown high-dimensional parameter of interest with $\theta_i = \mathbb{E}_f Y_i = \int_0^1 \varphi_i(x) f(x)dx$, $Y_i = \frac{1}{n}\sum_{l=1}^n \varphi_i(X_l)$, and $\sigma_n \xi_i = Y_i - \theta_i$. Since $|Y_i| = |\frac{1}{n}\sum_{l=1}^n \varphi_i(X_l)| \le c_\varphi$, we have $\sigma_n|\xi_i| \le |Y_i| + |\theta_i| \le 2c_\varphi$ and $\mathrm{Var}(\sigma_n\xi_i) \le \frac{c_\varphi^2}{n}$. The parameter σ_n will be chosen later, for now it is any sequence $\sigma_n \in [0,1]$.

Notice that we reduced the original density estimation problem to a finite dimensional version of the model from Section 5.5.1, however the errors ξ_i's are now not iid normals, which complicates the study of the present model. Consider the same smoothness structure (5.46) as in Section 5.5.1, with the difference that we restrict the family of structures $I \in \mathcal{I} = [n]_0$. The oracle rate becomes

$$r^2(\theta) = \min_{I\in[n]_0} \big(\textstyle\sum_{i\ge I+1}\theta_i^2 + \sigma_n^2 I\big) = \sum_{i\ge I_o+1}\theta_i^2 + \sigma_n^2 I_o.$$

Conditions (A2) and (A3) are met in the same way as for the signal+noise model from Section 5.5.1. However, in order to derive at least the local estimation and posterior contraction results, we also need Condition (A1). This condition is now not immediate since the errors ξ_i's are non-normal and dependent in the model (5.49) (actually, the ξ_i's are asymptotically normal, but we are not going to rely on this). We apply the following strategy: introduce certain event and establish that the probability of this event is exponentially small (in n); next, under this event establish Condition (A1); finally, combine these two facts to derive the local estimation and posterior contraction results.

The following proposition is a direct consequence of McDiarmid's inequality; see, for instance Theorem 6.2 in [24].

Proposition 5.1. *For any $t > 0$ and $i \in [n]$,*

$$\mathbb{P}(\sigma_n|\xi_i| \ge t) \le 2\exp\big\{-c_\varphi^{-2}t^2 n/2\big\}.$$

The relation $\mathbb{P}(\max_{i\in[n]}|\xi_i| \ge t) \le \sum_{i\in[n]}\mathbb{P}(|\xi_i| \ge t)$ and Proposition 5.1 imply that, for the event $E = \{\max_{i\in[n]}|\xi_i| \le \sqrt{2}c_\varphi\}$,

$$\mathbb{P}(E^c) = \mathbb{P}\big(\max_{i\in[n]}|\xi_i| > \sqrt{2}c_\varphi\big) \le 2\exp\{-n\sigma_n^2 + \log n\}. \tag{5.50}$$

Now, by using (5.50), we ensure Condition (A1) under the event $E = \{\max_{i\in[n]}|\xi_i| \le \sqrt{2}c_\varphi\}$ with $\alpha = 1 \wedge 1/(2c_\varphi^2)$. Exactly, for any $I \in [n]_0$,

$$\begin{aligned}\mathbb{E}\exp\big\{\alpha\|\mathrm{P}_I\xi\|^2\big\}1_E &= \mathbb{E}\exp\Big\{\alpha\sum_{i=1}^I \xi_i^2\Big\}1\{\max_{i\in[n]}|\xi_i| \le \sqrt{2}c_\varphi\}\\ &\le \exp\big\{\alpha 2c_\varphi^2 I\big\} = e^I = \exp\{d_{s(I)}\}.\end{aligned} \tag{5.51}$$

We have thus verified the conditional version of Condition (A1) (under event E) and Conditions (A2) and (A3) for the model (5.49). This means that we can derive results on estimation, posterior contraction and uncertainty quantification for the density f in terms of the model (5.49). These are the counterparts of claims (i)-(v) of Corollary 5.1 summarized by Theorem 5.5 below. To the best of our knowledge, local results on uncertainty quantification for the density are new. In the below theorem, we keep the same notation for all the quantities involved as in the general framework, with the understanding that these are specialized for the model (5.49) with the smoothness structure and the oracle rate $r^2(\theta)$.

Theorem 5.5. *Let the constants $M_0, M_1, M_3, H_0, H_1, H_2, H_3, m_0, m_1, m_2, m_3, c_2, c_3, C_v$ be defined in Theorems 5.1-5.3 and (5.50). Then for any $M \geq 0$,*

$$\sup_{\theta \in \ell_2} \mathbb{E}_\theta \hat{\pi}\big(\|\theta - \vartheta\|^2 \geq M_0 r^2(\theta) + M\sigma_n^2 | Y\big) \leq 2e^{-n\sigma_n^2 + \log n} + H_0 e^{-m_0 M},$$

$$\sup_{\theta \in \ell_2} \mathbb{P}_\theta\big(\|\hat{\theta} - \theta\|^2 \geq M_1 r^2(\theta) + M\sigma_n^2\big) \leq 2e^{-n\sigma_n^2 + \log n} + H_1 e^{-m_1 M},$$

$$\sup_{\theta \in \ell_2} \mathbb{E}_\theta \hat{\pi}\big(I : r^2(I, \theta) \geq c_3 r^2(\theta) + M\sigma_n^2 | Y\big) \leq 2e^{-n\sigma_n^2 + \log n} + C_v e^{-c_2 M},$$

$$\sup_{\theta \in \ell_2} \mathbb{P}_\theta\big(\hat{r}^2 \geq M_3 r^2(\theta) + (M+1)\sigma_n^2\big) \leq 2e^{-n\sigma_n^2 + \log n} + H_3 e^{-m_3 M},$$

$$\sup_{\theta \in \ell_2 \cap \Theta_{\mathrm{eb}}(t)} \mathbb{P}_\theta\big(\theta \notin B(\hat{\theta}, \hat{R}_M)\big) \leq 2e^{-n\sigma_n^2 + \log n} + H_2 e^{-m_2 M}.$$

Let us outline the idea of the proof (which is omitted) of the first claim of the above theorem; the same reasoning applies to the remaining claims. The expectation of the empirical Bayes posterior probability $\mathbb{E}_\theta \Pi = \mathbb{E}_\theta \hat{\pi}\big(\|\theta - \vartheta\|^2 \geq M_0 r^2(\theta) + M\sigma_n^2 | Y\big)$ is bounded by the sum of two terms $\mathbb{E}_\theta \Pi \leq \mathbb{P}_\theta(E^c) + \mathbb{E}_\theta \Pi 1_E$. The first term is evaluated by using (5.50) (obtaining the bound $2e^{-n\sigma_n^2 + \log n}$); the second term is evaluated exactly in the same way as in the proof Theorem 5.1, because Condition (A1) is fulfilled under the event E according to (5.51). Counterparts of assertions (ii) and (iii) of Theorem 5.2 can also be formulated and proved in the same way. Notice that the results that rely on Condition (A4) are not claimed as we are unable to verify this condition at the moment.

As to the choice of σ_n^2 in the oracle rate, clearly, we would want it to be as small as possible. On the other hand, we want the claims of the theorem to be non-void, which is ensured only if $\sigma_n^2 n \geq C \log n$, or $\sigma_n^2 \geq \frac{C \log n}{n}$, for sufficiently large $C > 0$. In the sequel we take therefore $\sigma_n^2 = \frac{C \log n}{n}$. An extra log factor thus appeared which will also enter the minimax rates in the global results. We conjecture that one can get rid of that factor by using more accurate concentration inequalities when establishing Condition (A1).

As usually, the local results of Theorem 5.5 will imply global minimax adaptive results simultaneously over all scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$ (i.e., for which (5.45) holds). Hence, the same adaptive minimax results for the same scales as in Section 5.5.1 follow, up to a log factor as we have $\sigma_n^2 \asymp \frac{\log n}{n}$ in the model (5.49) instead of n^{-1} in the model from Section 5.5.1. The reader is invited to formulate a number of local and adaptive minimax results for this case. We should mention that it seems possible to extend the results to other structures (e.g., sparsity) and scales (e.g., Besov scales).

5.5.6. BICLUSTERING MODEL

Suppose we observe a matrix $Y = (Y_{ij}) \in \mathbb{R}^{n_1 \times n_2}$:

$$Y_{ij} = \theta_{ij} + \sigma \xi_{ij}, \quad i = 1, \dots, n_1, \quad j = 1, \dots, n_2,$$

where $\theta = (\theta_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ is an unknown high-dimensional parameter of interest with *biclustering* structure, $\sigma > 0$ is the known noise intensity, $\xi = (\xi_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ is a random matrix with $\mathbb{E}_\theta \xi_{ij} = 0$. Without loss of generality, we set $\sigma = 1$ for the rest of this section. This model has been studied in detail in Chapter 4, here we consider it once again from

the perspective of general framework of projection structures. Some computations from Chapter 4 will be repeated for completeness.

The essence of biclustering structure is to reduce dimensionality of a large matrix of parameters by simultaneous grouping of the rows and columns. For example, if the rows of θ correspond to objects and the columns to features, a biclustering structure means that only a few features are relevant for identifying a few groups of similar objects. Biclustering structure means that the rows and columns of the matrix $\theta = (\theta_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ are split into k_1 and k_2 clusters, respectively, and the values θ_{ij} are the same for i, j from the same clusters. Let us give the mathematical formalization of this idea.

For $(k_1, k_2) \in [n_1] \times [n_2]$, consider a mapping $z = (z_1, z_2) : [n_1] \times [n_2] \mapsto [k_1] \times [k_2]$, where $z_1 : [n_1] \mapsto [k_1]$ and $z_2 : [n_2] \mapsto [k_2]$. Each mapping $z \in [k_1]^{[n_1]} \times [k_2]^{[n_2]}$ determines the pertinent partition $I = I(z)$ of the rows and columns of any matrix $(M_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ into $k_1 \times k_2$ blocks:

$$[n_1] \times [n_2] = z^{-1}([k_1] \times [k_2]) = z_1^{-1}([k_1]) \times z_2^{-1}([k_2]) = \cup_{(I_i^1, I_j^2) \in I} (I_i^1, I_j^2),$$

5

where $I_i^1 = z_1^{-1}(i)$ and $I_j^2 = z_2^{-1}(j)$. The *biclustering structure* is nothing else but just this partition $I = I(z) = (I^1, I^2)$, where $I^1 = I^1(z_1) = (I_i^1 : i \in [k_1])$ is the row partition and $I^2 = I^2(z_2) = (I_j^2 : j \in [k_2])$ is the column partition. So, the collection of all mappings $\mathcal{Z} = \mathcal{Z}(n_1, n_2) = \{(z_1, z_2) \in [k_1]^{[n_1]} \times [k_2]^{[n_2]}, (k_1, k_2) \in [n_1] \times [n_2]\}$ yields the collection of all biclustering structures (which are all *biclustered* partitions of $[n_1] \times [n_2]$):

$$\mathcal{I} = \mathcal{I}(n_1, n_2) = \big\{I(z), z \in [k_1]^{[n_1]} \times [k_2]^{[n_2]}, (k_1, k_2) \in [n_1] \times [n_2]\big\}.$$

A biclustering structure $I \in \mathcal{I}$ in terms of parameter θ is expressed by imposing $\theta \in \mathbb{L}_I \subseteq \mathbb{R}^{n_1 n_2}$, where the linear subspace $\mathbb{L}_I$ is defined as

$$\mathbb{L}_I = \big\{x \in \mathbb{R}^{n_1 n_2} : x_{ij} = x_{i'j'} \ \forall\ (i, j), (i', j') \in (I_1, I_2),\ \forall\ (I_1, I_2) \in I\big\}. \tag{5.52}$$

Assume that $\mathcal{I}$ is "cleaned up" in the sense that $\mathbb{L}_I \neq \mathbb{L}_{I'}$ for all $I \neq I'$ (see Remark 1.1).

The structural slicing mapping $s : \mathcal{I} \mapsto \mathcal{S}$ is defined as $s(I) = (s_1(I), s_2(I)) \in [n_1] \times [n_2] \triangleq \mathcal{S}$, where $(s_1(I), s_2(I))$ denotes the numbers of nonempty row and column blocks in the structure $I \in \mathcal{I}$. Then $d_{s(I)} = d_I = \dim(\mathbb{L}_I) = s_1(I)s_2(I)$.

Let us propose a reasonable majorant $\rho(s)$ for the layer complexity $d_s + \log|\mathcal{I}_s| = s_1 s_2 + \log|\mathcal{I}_s|$. Clearly, $|\mathcal{I}_s| \le N(n_1, s_1)N(n_2, s_2)$, where $N(n, k)$ is the number of ways to put n different objects into k different boxes so that each box contains at least one object. Notice that $S(n, k) = N(n, k)/k! = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$ is a Stirling number of the second kind. To have a simple closed form expression for a majorant of the complexity, instead of $N(n_1, s_1)N(n_2, s_2)$ we can use its upper bound $s_1^{n_1}s_2^{n_2}$ (all the partitions of $[n_1] \times [n_2]$ into $s_1 \times s_2$ blocks, some of which are possibly empty). However, the bound $|\mathcal{I}_s| \le s_1^{n_1}s_2^{n_2}$ becomes too crude for some $s \in \mathcal{S}$. In particular, this bound is too crude for the cases (i) $(s_1, s_2) \in \mathcal{S}_1 = \{(s_1, s_2) \in [n_1] \times [n_2] : s_1 < n_1, s_2 = n_2\}$, (ii) $(s_1, s_2) \in \mathcal{S}_2 = \{(s_1, s_2) \in [n_1] \times [n_2] : s_1 = n_1, s_2 < n_2\}$, and (iii) $(s_1, s_2) \in \mathcal{S}_3 = \{(n_1, n_2)\}$. Indeed, let $\mathrm{id}_m : [m] \mapsto [m]$ with $\mathrm{id}_m(s) = s$, $s \in [m]$, the identity mapping of $[m]$. Then it is easy to see that $\mathbb{L}_{I(z_1,z_2)} = \mathbb{L}_{I(z_1,\mathrm{id}_{n_2})}$ for all $z_2 \in [n_2]^{[n_2]}$ and all $z_1 \in [s_1]^{[n_1]}$, $s_1 \in [n_1]$. Similarly, $\mathbb{L}_{I(z_1,z_2)} = \mathbb{L}_{I(\mathrm{id}_{n_1},z_2)}$ for all $z_1 \in [n_1]^{[n_1]}$, $z_2 \in [s_2]^{[n_2]}$, $s_2 \in [n_2]$; and

$\mathbb{L}_{I(z_1,z_2)} = \mathbb{L}_{I(\mathrm{id}_{n_1},\mathrm{id}_{n_2})}$ for all $z_1 \in [n_1]^{n_1}, z_2 \in [n_2]^{n_2}$. Hence, $|\mathcal{I}_s| \le \left|[s_1]^{[n_2]}\right| \le s_1^{n_1}$ for $(s_1,s_2) \in \mathcal{S}_1$, $|\mathcal{I}_s| \le s_2^{n_2}$ for $(s_1,s_2) \in \mathcal{S}_2$, and $|\mathcal{I}_s| \le 1$ for $(s_1,s_2) \in \mathcal{S}_3$. Thus, we improve the bound $d_s + \log|\mathcal{I}_s| \le s_1 s_2 + \log(s_1^{n_1} s_2^{n_2})$ by proposing the following majorant $\rho(s)$ for the complexity of the layer $\mathcal{I}_s$: $d_s + \log|\mathcal{I}_s| \le \rho(s)$, where $\rho(s)$ is defined as follows:

$$\rho(s) \triangleq \begin{cases} s_1 s_2 + n_1 \log s_1 + n_2 \log s_2, & s_1 < n_1,\ s_2 < n_2, \\ s_1 n_2 + n_1 \log s_1, & s_1 < n_1,\ s_2 = n_2, \\ n_1 s_2 + n_2 \log s_2, & s_1 = n_1,\ s_2 < n_2, \\ n_1 n_2, & s_1 = n_1,\ s_2 = n_2. \end{cases} \tag{5.53}$$

This is an example of the so called *elbow effect* mentioned in Remark 5.3.

In case $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, Conditions (A1) and (A4) hold with $d_{s(I)} = \dim(\mathbb{L}_I)$ in view of Remarks 1.6 and 5.10. In Section 4.1.4 of Chapter 4 it was shown that Condition (A1) (which is Condition (C1) in Chapter 4) is also fulfilled in case $Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$.

Let us verify Condition (A2): for any $\nu \ge 1$,

$$\sum_{I \in \mathcal{I}} e^{-\nu\rho(s(I))} \le \sum_{(s_1,s_2) \in [n_1]\times[n_2]} e^{-\nu s_1 s_2} = (e^{\nu} + e^{-\nu} - 2)^{-1} = C_\nu.$$

Thus, the properties (i)-(iv) of Corollary 5.1 follow for the biclustering model with the ξ_i' that are independent and either normal or binomial, in fact, for any ξ satisfying Condition (A1).

One can also check Condition (A3), so that the coverage property (v) of Corollary 5.1 holds under EBR as well. However, the peculiarity of the biclustering structure is that the size and coverage claims (vi)-(vii) for the confidence ball $B(\hat{\theta}, \tilde{R}_M)$ are stronger and more useful in this case than the corresponding claims (iv)-(v) for the confidence ball $B(\hat{\theta}, \hat{R}_M)$.

Indeed, the coverage property (v) holds uniformly only under the EBR, whereas the coverage property (vii) is uniform over the entire space $\Theta = \mathbb{R}^{n_1 \times n_2}$. So, basically the deceptiveness issue is not present in the coverage property (vii) for the confidence ball $B(\hat{\theta}, \tilde{R}_M)$, it appears only marginally in the size relation (vi) of Corollary 5.1. Indeed, the size $\tilde{R}_M$ of the ball is of the oracle rate order uniformly in $\theta \in \Theta \backslash \tilde{\Theta} = \mathbb{R}^{n_1 \times n_2} \backslash \tilde{\Theta}$, where $\tilde{\Theta}$ is defined by (5.28). By the definition of $\tilde{\Theta}$, $r^2(\theta) \ge c\sigma^2\sqrt{n_1 n_2}$ for $\theta \in \Theta \backslash \tilde{\Theta}$. For the biclustering model, we can take $c = \log 2$, and $\tilde{\Theta}$ can be written as $\tilde{\Theta} = \{\theta \in \mathbb{R}^{n_1 \times n_2} : \min\{s_{o1}(\theta), s_{o2}(\theta)\} = 1\}$ with $(s_{o1}(\theta), s_{o2}(\theta)) = s(I_o(\theta))$, where the oracle $I_o(\theta)$ is defined by (5.15). Hence, for the biclustering model, $\tilde{\Theta}$ is indeed a "thin" subset of $\mathbb{R}^{n_1 \times n_2}$ consisting of *highly structured parameters*, whose oracle number of either row or block columns is 1. As we have already discussed at the end of Section 5.2.4, this means that, modulo highly structured parameters, there is no deceptiveness phenomenon in the biclustering model.

Consider an example of scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the local rate $r^2(\theta)$.

Minimax results for the biclustering model. In [40], classes $\Theta^{\text{asym}}_{k_1 k_2}$ are introduced (and classes $\Theta_{k_1 k_2}(M)$ from [41]). In our notation, $\Theta^{\text{asym}}_{s_1 s_2} = \cup_{I \in \mathcal{I}_s} \Theta_I$, where $s = (s_1, s_2) \in [n_1] \times [n_2] \triangleq \mathcal{S}$, $\Theta_I \triangleq \mathbb{L}_I \cap [0,1]^{n_1 \times n_2}$ and $\mathbb{L}_I$ is defined by (5.52). So, the family of classes $\Theta^{\text{asym}}_{s_1 s_2}$ is

nothing else but the scale $\{\Theta_s, s \in \mathcal{S}\}$. The minimax rate $r^2(\Theta_s) \triangleq s_1 s_2 + n_1 \log s_1 + n_2 \log s_2$ over Θ_s is derived in [40], under the assumption $\log s_1 \asymp \log s_2$. It is easy to see that the oracle rate $r^2(\theta)$ covers the scale $\{\Theta_s, s \in \mathcal{S}\}$ in the sense of (5.45). Indeed, if $\theta \in \Theta_s$, then $\theta \in \mathbb{L}_{I'}$ for some $I' \in \mathcal{I}_s$, so that $\mathrm{P}_{I'}\theta = \theta$ and hence

$$r^2(\theta) \le r^2(I', \theta) = \rho(s(I')) = \rho(s) \le r^2(\Theta_s), \quad \theta \in \Theta_s. \tag{5.54}$$

Corollary 5.2 follows for this case with the minimax rate $r^2(\Theta_s)$ defined above.

According to Remark 5.16 we can derive the following upper bound for the oracle rate:

$$r^2(\theta) \le \min\{r^2(\Theta_s), s_1 n_2 + n_1 \log s_1, n_1 s_2 + n_2 \log s_2, n_1 n_2\} \triangleq \bar{r}^2(\Theta_s).$$

Remark 5.16. *As we have mentioned several times, in view of Remark 5.10, Condition (A4) is always fulfilled whenever $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. However, for the biclustering model (and the stochastic block model, described below), a more appropriate distribution for the observations is binomial, i.e., $Y_{ij} \overset{\text{ind}}{\sim}$ Bernoulli(θ_{ij}). This case is important in relation to network modeling. Unfortunately, we were unable to establish Condition (A4) for the binomial case. To be precise, verification of Condition (A4) for the binomial observations essentially boils down to the problem of estimating the functional $F(\theta) = \sum_{i,j} \theta_{ij}^2$ with the rate $\sqrt{N} = \sqrt{n_1 n_2}$. It is not known to us whether it is possible to construct such an estimator (possibly exploiting the biclustering structure of θ), which is an interesting and challenging problem on its own.*

STOCHASTIC BLOCK MODEL

Here we briefly discuss a particular case of biclustering model, the *stochastic block model* (SBM) which is used in the literature on networks to model undirected network graphs. Oracle estimation and posterior contraction rate results for stochastic block model were recently derived in [42] and [54]. Precisely, to get the SBM from the biclustering model, we assume additionally $s_1 = s_2 = s$, $n_1 = n_2 = n$, $z_1 = z_2 = z$. For a mapping $z \in [s]^{[n]}$, the pertinent row partition in the SBM is $I = I(z) = (z^{-1}(i), i \in [s])$, which is the same as the column partition.

In the binomial case $Y_{ij} \overset{\text{ind}}{\sim}$ Bernoulli(θ_{ij}), the observations Y_{ij} can be associated with network data. In this case Y_{ij} stands for the presence or absence of an edge between vertices i and j in the network interpretation. To model undirected network graphs, some conditions (called *network conditions*) are then additionally assumed: the "no self-loop" condition $Y_{ii} = \theta_{ii} = 0$ and symmetry condition $Y_{ij} = Y_{ji}$ and $\theta_{ij} = \theta_{ji}$. Denote by Θ_{net} the parameters $\theta \in \mathbb{R}^{n_1 \times n_2}$ satisfying these additional network conditions.

All the quantities, conditions and claims specialize to the SBM by setting $s_1 = s_2 = s$, $n_1 = n_2 = n$, $z_1 = z_2 = z$ in all the above formulas for the biclustering model. The linear subspaces $\mathbb{L}_I$ defined by (5.52) will get adjusted since $z_1 = z_2$, the family $\mathcal{I}_s$ can be associated with the collection of all possible partitions of $[n]$ into s blocks, parametrized by mappings $z \in [s]^{[n]}$. $|\mathcal{I}_s| \le s^n$, $s \in \mathcal{S} \triangleq [n]$. The structural slicing mapping $s(I)$ is the number of blocks in the partition I. Notice that under additional network conditions $cs^2(I) \le \dim(\mathbb{L}_I) \le s^2(I)$, so that we can use $s^2(I)$ (instead of the true $d_I = \dim(\mathbb{L}_I)$) in the complexity part of the local rate as it is still of the same order, although some constants

can be improved because of this extra network structure. We have $d_{s(I)} = d_I = \dim(\mathbb{L}_I) \le s^2(I)$, $\log|\mathcal{I}_s| \le n\log s$, and we take $\rho(s(I)) = s^2(I) + n\log s(I)$. Conditions (A1)-(A4) are fulfilled in the same way as for the biclustering model, leading to Corollary 5.1. As to the binomial case, see Remark 5.16.

Consider a couple of examples of scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the local rate $r^2(\theta)$.

Minimax results for the stochastic block model. In [40] (cf. [54]), classes Θ_k were introduced. In our notation, $\Theta_s = \cup_{I\in\mathcal{I}_s}\Theta_I$, where $s = k$, $\Theta_I = \mathbb{L}_I \cap \Theta_{\rm net} \cap [0,1]^{n^2}$, $I \in \mathcal{I}_s$, $s \in \mathcal{S}$. So, we have the scale $\{\Theta_s, s \in \mathcal{S}\}$ and the adaptive minimax results over this scale follow from the local results given by Corollary 5.1. Indeed, as is shown in [40], the minimax rate over Θ_s in the SBM is $r^2(\Theta_s) = \inf_{\hat\theta}\sup_{\theta\in\Theta_s}\mathbb{E}_\theta\|\hat\theta-\theta\|^2 \asymp s^2 + n\log s = \rho(s)$. On the other hand, for each $\theta \in \Theta_s$ there exists $I \in \mathcal{I}_s$ such that $\theta \in \mathbb{L}_I$. Hence, $\mathrm{P}_I\theta = \theta$ and $r^2(\theta) \le r^2(I,\theta) = \rho(s) \asymp r^2(\Theta_s)$. This implies Corollary 5.2 for this scale.

Remark 5.17. *As to the deceptiveness phenomenon in the SBM, for the confidence ball $B(\hat\theta, \tilde R_M)$ we again have the coverage property uniformly over the whole scale $\{\Theta_s, s \in \mathcal{S}\}$, whereas the size property with the optimal radial rate holds over all classes $\{\Theta_s, s = 2,\dots,n\}$, but one: Θ_1. Indeed, the class Θ_1 consists of highly structured parameters $\theta \in \mathbb{R}^{n^2}$, whose coordinates are all equal. The case $\theta \in \Theta_1$ reduces to just one-dimensional signal+noise model with $N = n^2$ observations. Since the effective radial rate $g_M(\theta, N)$ for the confidence ball $B(\hat\theta, \tilde R_M)$ is always at least of the order $n\sigma^2 \gg \sigma^2 = r^2(\Theta_1)$, we could not attain the optimal rate $r^2(\Theta_1)$ in the size relation only for the highly structured parameters $\theta \in \Theta_1$.*

Graphon classes. It is also possible to derive the global minimax results for the function class of graphons as consequence of our local results. In Section 4.3.2 the relation (5.45) was established for the function class of graphons, so that Corollary 5.2 follows with corresponding minimax rate of the function class of graphons, defined in Section 4.3.2.

5.5.7. LINEAR REGRESSION

Let us consider a general setting for linear regression:

$$Y = \mathrm{X}\beta + \sigma\xi, \tag{5.55}$$

where $\mathrm{X} = \mathrm{diag}(\mathrm{X}^1,\dots,\mathrm{X}^m) \in \mathbb{R}^{mn\times mp}$ is a block diagonal matrix, whose blocks $\mathrm{X}^1,\dots,\mathrm{X}^m \in \mathbb{R}^{n\times p}$ are design matrices, $\sigma > 0$ is the known noise intensity, $\beta = (\beta^1,\dots,\beta^m) \in \mathbb{R}^{mp}$ is a concatenation of m unknown p-dimensional vectors $\beta^1,\dots,\beta^m \in \mathbb{R}^p$, $Y = (Y^1,\dots,Y^m) \in \mathbb{R}^{mn}$ is a concatenation of observed vectors $Y^1,\dots,Y^m \in \mathbb{R}^n$, $\xi = (\xi_i, i \in [mn])$, $\xi_i \overset{\rm ind}{\sim} \mathrm{N}(0,1)$.

Wherever appropriate, by M_I we denote the submatrix of M with columns $(M_i, i \in I)$, x_I is the $|I|$-dimensional subvector of $x \in \mathbb{R}^p$ with coordinates $i \in I$, $\|\beta\|_0$ denotes the number of non-zero elements of β, i.e., the cardinality of the support $I^*(\beta) = \mathrm{supp}(\beta) = \{i : \beta_i \ne 0\}$ of β. Under $\xi_i \overset{\rm ind}{\sim} \mathrm{N}(0,1)$, Conditions (A1) and (A4) will hold with $d_{s(I)} = \dim(\mathbb{L}_I)$ for all cases of this section in view of Remarks 1.6 and 5.10.

We should emphasize that in this section we denote by β the vector of unknown parameters in (5.55), notation commonly used in the literature. This is not to be confused with the structural parameter β for indexing the scales of classes $\{\Theta_\beta, \beta \in \mathcal{B}\}$ which we use in other sections, in this section we will instead use the notation $\{\Theta_\gamma, \gamma \in \Gamma\}$ for scales.

Many particular linear models can be put in (5.55) by choosing appropriately m, p and X. The case $m = 1, p = n, \mathrm{X} = \mathrm{I}$ is already considered in Section 5.5.1 for the smoothness structure and in Section 5.5.3 for the sparsity structure. In the following subsections we consider several other specific models and structures in detail.

Remark 5.18. *The true model does not have to be exactly of the form* (5.55)*, it can be* $Y = \theta + \sigma\xi$*, where* $\theta \neq \mathrm{X}\beta$*. In that case,* (5.55) *is an approximating model of the true model and all the local results hold with* θ *substituted everywhere instead of* $\mathrm{X}\beta$*. The global minimax results will have to be modified by including the approximation term* $\sup_{\theta\in\Theta_\gamma} \mathbb{E}_\theta \|\theta - \mathrm{P}_{I_o}\theta\|^2$ *in the minimax rate* $r^2(\Theta_\gamma)$*.*

LINEAR REGRESSION: $m = 1$

This is the classical linear regression model, and in the high-dimensional setting typically $p \gg n$. So, to be able to make sensible inference, one needs to exploit a structure on β. The vector $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ with a structure $I \subseteq [p]$ is assumed to be *sparse* in the sense that $\beta_i = 0$ for $i \in I^c$, that is, the predictors $(X_i : i \in I^c)$ of design matrix $\mathrm{X} = (X_1, \dots, X_p)$ are irrelevant from the perspective of structure I. Denote $r = r(\mathrm{X}) = \mathrm{rank}(\mathrm{X})$.

A local approach for this model, delivering also the adaptive minimax results for many sparsity structures simultaneously, is considered in [42] and [11] for the posterior contraction rates. In [11] also the uncertainty quantification is treated. Note that "signal+noise" model is a special case of linear regression with $\mathrm{X} = \mathrm{I}$ and $p = n$, and it has been treated in several previous sections.

In this model, the sparsity structure is expressed by the linear spaces

$$\mathbb{L}_I = \{\mathrm{X}_I x_I \in \mathbb{R}^n : x_I \in \mathbb{R}^{|I|}\} = \{\mathrm{X}x \in \mathbb{R}^n : x \in \mathbb{R}^p,\ x_i = 0 \text{ for } i \notin I\}, \tag{5.56}$$

$I \in \mathcal{I}$, where the family of structures is $\mathcal{I} = \mathcal{I}_1 \cup \{I_r\}$ with $\mathcal{I}_1 = \{I \subseteq [p] : 2|I|\log(ep/|I|) \le r\}$ (recall that $r = \mathrm{rank}(\mathrm{X})$) and $I_r = (i_1, \dots, i_r) \subseteq [p]$ such that $(X_{i_1}, \dots, X_{i_r})$ are r linearly independent columns of X. Then $|\mathcal{I}| \le 2^p$, the structural slicing mapping is taken to be $s(I) = |I| \in \mathcal{S} \triangleq [r]_0$. Further, we have $\theta = \mathrm{X}\beta$, $d_I = \dim(\mathbb{L}_I) \le \min\{|I|, r\} \le \min\{|I|, n, p\}$ for $I \in \mathcal{I}_1$ and $d_{I_r} = \dim(\mathbb{L}_{I_r}) = r$. Clearly, $\log|\mathcal{I}_{s(I)}| = \log\binom{p}{|I|} \le |I|\log(\frac{ep}{|I|})$ for $I \in \mathcal{I}_1$ and $\log|\mathcal{I}_{s(I_r)}| = 0$. Since $d_I + \log|\mathcal{I}_{s(I)}| \le |I| + |I|\log(\frac{ep}{|I|}) \le 2|I|\log(\frac{ep}{|I|})$ for $I \in \mathcal{I}_1$ and $d_{s(I_r)} + \log|\mathcal{I}_{s(I_r)}| = d_{I_r} = r$, we take the majorant

$$\rho(s(I)) = 2|I|\log(\tfrac{ep}{|I|})1\{I \in \mathcal{I}_1\} + r1\{I = I_r\}, \quad I \in \mathcal{I}. \tag{5.57}$$

Notice that we could use a smaller majorant $\rho'(s(I)) = d_I + \log\binom{p}{|I|}$ for $I \in \mathcal{I}_1$ (the best choice), but this majorant is not practical to use.

Remark 5.19. *In the majorant* $\rho(s(I))$ *defined above, we see the* elbow effect *mentioned in Remark 5.3, this elbow effect will enter the rate as well. Let us explain how this elbow effect has emerged in this model.*

Notice that we could consider the more natural full family of structures $\bar{\mathcal{I}} = \{J : J \subseteq [p]\}$, *so that* $|\bar{\mathcal{I}}| = 2^p$, *with the same structural slicing mapping* $s(I) = |I| \in \mathcal{S} \triangleq [p]_0$, *but defined on the family* $\bar{\mathcal{I}}$. *As before,* $d_I = \dim(\mathbb{L}_I) \le \min\{|I|, r\}$ *and* $|\mathcal{I}_{s(I)}| = \binom{p}{|I|}$. *Since* $d_I + \log|\mathcal{I}_{s(I)}| \le |I| + |I|\log(\frac{ep}{|I|}) \le 2|I|\log(\frac{ep}{|I|})$, *the majorant would be* $\bar{\rho}(s(I)) = 2|I|\log(\frac{ep}{|I|})$, $I \in \bar{\mathcal{I}}$. *The idea of the family* $\mathcal{I}$ *is that, even though* $\mathcal{I} \subseteq \bar{\mathcal{I}}$, *the family* $\mathcal{I}$ *still covers* $\bar{\mathcal{I}}$ *in the sense of Remark 5.3. Indeed,* $r^2(I,\beta) = \|\mathrm{X}\beta - \mathrm{P}_I\mathrm{X}\beta\|^2 + \sigma^2\rho(s(I)) = \|\mathrm{X}\beta - \mathrm{P}_I\mathrm{X}\beta\|^2 + \sigma^2\bar{\rho}(s(I)) = \bar{r}^2(I,\beta)$ *for* $I \in \mathcal{I}_1$, *and* $r^2(I_r,\beta) = \sigma^2 r \le \sigma^2 2|I|\log(\frac{ep}{|I|}) \le \bar{r}^2(I,\beta)$ *for all* $I \in \bar{\mathcal{I}}\backslash\mathcal{I}_1$, *as* $\mathrm{P}_{I_r}\mathrm{X}\beta = \mathrm{X}\beta$ *for any* $\beta \in \mathbb{R}^p$.

Here we considered an important case when a seemingly right (full) family $\bar{\mathcal{I}}$ *of structures can be reduced to a subfamily* $\mathcal{I} \subset \bar{\mathcal{I}}$ *that has a reduced complexity but still covers the original family* $\bar{\mathcal{I}}$ *in the sense of Remark 5.3, thus improving the resulting oracle rate. This is a typical situation exhibiting the "elbow effect" in the complexity term of the rate; below there are a couple of more such example (also Section 5.5.10).*

Remark 5.20. *We could further reduce the family of structures to* $\mathcal{I}' = \{I_r\} \cup \mathcal{I}'_1$, *with* $\mathcal{I}'_1 = \{I \subseteq \mathcal{I}_1 :$ *the columns* $(X_i, i \in I)$ *are linearly independent*$\}$ *(with the same structural slicing mapping* $s(I) = |I|$*), so that* $\mathcal{I}' \subseteq \mathcal{I}$. *In this case, we have* $d_I = \dim(\mathbb{L}_I) = |I|$, $|I| \le r \le \min\{n, p\}$ *for each* $I \in \mathcal{I}'$, *the layer is* $\mathcal{I}_{s(I)} = \{J \subseteq \mathcal{I}_1 : \dim(\mathbb{L}_J) = \dim(\mathbb{L}_I)\}$ *for* $I \in \mathcal{I}'_1$ *and* $\mathcal{I}_{s(I_r)} = \{I_r\}$. *In this case, we can take the majorant* $\bar{\rho}(s(I)) = (|I| + \log|\mathcal{I}_{s(I)}|)1\{I \in \mathcal{I}_1\} + r1\{I = I_r\}$. *When implementing the Bayesian or penalization procedure, the majorant* $\rho(s(I)) = 2|I|\log(ep/|I|)1\{I \in \mathcal{I}_1\} + r1\{I = I_r\}$ *is more practical to use also for the family* $\mathcal{I}'$. *But then* $\mathcal{I}$ *also covers* $\mathcal{I}'$, *thus boiling down to the same resulting oracle rate. Therefore, as soon as we use the same majorant, it does not matter which family of structures,* $\mathcal{I}$ *or* $\mathcal{I}'$, *we take. A slightly bigger constant in Condition (A2) will be for the family* $\mathcal{I}$ *as there are more terms in the sum. We will use the family* $\mathcal{I}$.

Condition (A2) is fulfilled, since, according to Remark 1.8, for any $\nu > 1$

$$\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} \le \sum_{s\in\mathcal{S}} e^{-(\nu-1)\rho(s(I))} \le \sum_{s=0}^{p} e^{-(\nu-1)s} \le \frac{1}{1-e^{1-\nu}} = C_\nu.$$

As to Condition (A3), for any $I_0, I_1 \in \mathcal{I}$, take $I' = I_r$ if either $I_0 = I_r$ or $I_1 = I_r$ or $2|I_0 \cup I_1|\log(ep/|I_0 \cup I_1|) > r$; otherwise take $I' = I_0 \cup I_1$. Since $(\mathbb{L}_{I_0} \cup \mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'} = \mathbb{L}_{I_0} + \mathbb{L}_{I_1}$ and $\rho(s(I')) \le \rho(s(I_0)) + \rho(s(I_1))$, Condition (A3) is also fulfilled.

As consequence of our general results, we obtain Corollary 5.1 with the local (prediction) rate $r^2(\beta) = \min_{I\in\mathcal{I}} r^2(I,\beta) = \min_{I\in\mathcal{I}}\{\|\mathrm{X}\beta - \mathrm{P}_I\mathrm{X}\beta\|^2 + \sigma^2\rho(s(I))\}$, where the majorant $\rho(s(I))$ is defined by (5.57). In particular,

$$\begin{aligned} r^2(\beta) &\le r^2(I^*(\beta),\beta) \wedge r^2(I_r,\beta) = \sigma^2\big[\rho(s(I^*(\beta))) \wedge \rho(s(I_r))\big] \\ &\lesssim \sigma^2\big[\big(|I^*(\beta)|\log(\tfrac{ep}{|I^*(\beta)|})\big) \wedge r\big]. \end{aligned} \tag{5.58}$$

Notice that the claims (i)-(vii) of Corollary 5.1 deliver finer and stronger versions of the corresponding results from [11]; besides, we can drop the normality and independence assumptions and impose only Condition (A1) instead.

Next, by virtue of Corollary 5.2 the local results imply global minimax adaptive results at once over all scales $\{\Theta_\gamma, \gamma \in \Gamma\}$ covered by the oracle rate $r^2(\beta)$ (i.e., for which (5.45) holds). Below we present a couple of scales $\{\Theta_\gamma, \gamma \in \Gamma\}$ covered by the oracle rate $r^2(\beta)$.

Nearly black vectors. For $s \in [p]$, introduce

$$\ell_0[s] = \{\beta \in \mathbb{R}^p : \|\beta\|_0 = |I^*(\beta)| \le s\}, \text{ where } I^*(\beta) = \{i \in [p] : \beta_i \neq 0\},$$

the set of vectors with at most s nonzero elements. Under certain conditions on the parameters s, p, n and the design matrix X (at least, $s\log(ep/s) \lesssim r = \mathrm{rank}(\mathrm{X})$ has to hold), the minimax prediction estimation rate over $\ell_0[s]$ is known to be $r^2(\ell_0[s]) = \inf_{\hat\beta}\sup_{\beta\in\ell_0[s]}\mathbb{E}_\beta\|\mathrm{X}\hat\beta - \mathrm{X}\beta\|^2 \asymp \sigma^2 s\log(ep/s)$; see [27] and [73]. The adaptive minimax results for ℓ_0-balls were considered by [32], [42], [63] for posterior contraction rates and by [67] for uncertainty quantification problem.

If $\beta \in \ell_0[s]$, then $\mathrm{X}\beta \in \mathbb{L}_{I^*(\beta)}$ for $I^*(\beta) \in \mathcal{I}_s$ such that $\mathrm{P}_{I^*(\beta)}\mathrm{X}\beta = \mathrm{X}\beta$ and $|I^*(\beta)| \le s$, and hence $r^2(I^*(\beta), \beta) \le \sigma^2|I^*(\beta)|\log\big(ep/|I^*(\beta)|\big) \le \sigma^2 s\log\big(ep/s\big)$. By the definition (5.15) of the oracle rate $r^2(\beta)$, we have that $r^2(\beta) \le r^2(I^*(\beta), \beta) \wedge r^2(I_r, \beta)$. Then we obtain trivially that

$$\sup_{\beta\in\ell_0[s]} r^2(\beta) \le \sup_{\beta\in\ell_0[s]} \big[r^2(I^*(\beta),\beta) \wedge r^2(I_r,\beta)\big] \lesssim \sigma^2\big[r \wedge \big(s\log(\tfrac{ep}{s})\big)\big] \lesssim r^2(\ell_0[s]).$$

We thus established the relation (5.45) for the scale $\ell_0[s]$, and Corollary 5.2 follows with the minimax rate $r^2(\ell_0[s])$ defined above.

Weak ℓ_q-balls. For $q \in (0, 1]$, the *weak ℓ_q-ball* is defined by

$$\ell_q[R] = \big\{\beta \in \mathbb{R}^p : \beta^2_{[i]} \le R^2 i^{-2/q},\ i \in [p]\big\},\ R^2 \ge \sigma^2 \log p,$$

where $\beta^2_{[1]} \ge \ldots \ge \beta^2_{[p]}$ are the ordered $\beta_1^2, \ldots, \beta_p^2$. We assume that there exists a constant $L > 0$ such that $\max_{i\in[p]} \|X_i\|^2 \le nL^2$. The minimax prediction estimation rate over $\ell_q[R]$ in ℓ_2-prediction norm is known to be $r^2(\ell_q[R]) = \inf_{\hat\beta}\sup_{\beta\in\ell_q[R]}\mathbb{E}_\beta\|\mathrm{X}\hat\beta - \mathrm{X}\beta\|^2 = R^q n^{q/2}\sigma^{2-q}[\log(1 + \frac{p\sigma^q}{n^{q/2}R^q})]^{1-q/2}$ when $R^2 \ge \sigma^2\log p$; see [84] (cf. [38], [23]). The adaptive minimax results for *weak* ℓ_q-balls in ℓ_2-prediction norm were considered by [42] for posterior contraction rates.

Define $j = O_\beta(i)$ if $\beta_i^2 = \beta^2_{[j]}$, with the convention that in the case $\beta^2_{i_1} = \ldots = \beta^2_{i_k}$ for $i_1 < \ldots < i_k$ we set $O_\beta(i_{l+1}) = O_\beta(i_l) + 1$, $l = 1, \ldots, k-1$. Let $I^* = I^*(\beta) = \{i \in [p] : O_\beta(i) \le R^*\}$ with $R^* = e(\frac{R}{\sigma})^q n^{q/2}[\log(\frac{p\sigma^q}{n^{q/2}R^q})]^{-q/2}$, and $\beta^* = \beta^*(\beta) = ((\beta^*_i)_{i\in[p]} : \beta^*_i = \beta_i$ for $i \in I^*, \beta^*_j = 0$ for $j \notin I^*)$.

There exists $I^* \in \mathcal{I}$ such that $\mathrm{X}\beta^* \in \mathbb{L}_{I^*}$ and $\|\mathrm{X}\beta - \mathrm{P}_{I^*}\mathrm{X}\beta\|^2 = \|\mathrm{X}\beta - \mathrm{X}\beta^*\|^2$. By using this and the fact that $\max_{i\in[p]}\|X_i\|^2 \le L^2 n$, we derive (5.45):

$$\begin{aligned}\sup_{\beta\in\ell_q[R]} r^2(\beta) &\le \sup_{\beta\in\ell_q[R]} r^2(I^*,\beta) \le \sigma^2 R^* \log(\tfrac{ep}{R^*}) + \sup_{\beta\in\ell_q[R]}\|\mathrm{X}\beta - \mathrm{X}\beta^*\|^2 \\ &\lesssim \sigma^2 R^*\log(\tfrac{p\sigma^q}{n^{q/2}R^q}) + L^2 n R^2 (R^*)^{1-2/q} \\ &\lesssim R^q n^{q/2}\sigma^{2-q}\big[\log(1 + \tfrac{p\sigma^q}{n^{q/2}R^q})\big]^{1-q/2} \asymp r^2(\ell_q[R]).\end{aligned}$$

Corollary 5.2 follows for this case with the minimax rate $r^2(\ell_q[R])$ defined above.

Model selection. Besides inference on $\theta = \mathrm{X}\beta$, several interesting corollaries were established in [11] and they follow from our results exactly in the same way, we provide them here for completeness.

The first corollary concerns a bound on the size of the selected model. Similar to [32] and [63], the following assertion shows that the models with substantially higher size than the true one are unlikely according to the posterior $\hat{\pi}(I|Y)$ (which is in essence the penalization method in case $\hat{\pi}(I|Y) = \check{\pi}(I|Y)$).

Proposition 5.2. *Under the conditions of Corollary 5.1, for sufficiently large M_0'*

$$\sup_{\beta\in\mathbb{R}^p} \mathbb{E}_\beta \hat{\pi}(I : |I| > C_0\|\beta\|_0 | Y) \le C_v \exp\big\{-c_2\big(\tfrac{M_0'}{2} - c_3\big)\|\beta\|_0 \log\big(\tfrac{ep}{\|\beta\|_0}\big)\big\},$$

where $C_0 = \max\{M_0', {M_0'}^2/e\}$.

Proof. Note that for any $M_0' > 2c_3$, $|I| \ge M_0'\|\beta\|_0$ implies that

$$r^2(I,\beta) \ge \sigma^2|I|\log\big(\tfrac{en}{|I|}\big) \ge M_0'\sigma^2\|\beta\|_0 \log\big(\tfrac{en}{M_0'\|\beta\|_0}\big) \ge \tfrac{M_0'}{2}\sigma^2\|\beta\|_0\log\big(\tfrac{en}{\|\beta\|_0}\big),$$

provided $\|\beta\|_0 < en/{M_0'}^2$. Since $r^2(\beta) \le r^2(I^*(\beta),\beta) \le \sigma^2\|\beta\|_0\log(en/\|\beta\|_0)$, the above display implies that $r^2(I,\beta) \ge c_3 r^2(\beta) + M_0''\sigma^2$, where $M_0'' = (M_0'/2 - c_3)\|\beta\|_0\log(en/\|\beta\|_0)$. By (i) of Theorem 5.2 (M_0'' corresponds to M from Theorem 5.2), the assertion holds for any $|I| \ge M_0'\|\beta\|_0$ whenever $\|\beta\|_0 < en/{M_0'}^2$. If $\|\beta\|_0 \ge en/{M_0'}^2$, the result trivially holds for any $|I| \ge {M_0'}^2\|\beta\|_0/e$. Hence, the choice $C_0 = \max\{M_0', {M_0'}^2/e\}$ ensures the result for any $\beta \in \mathbb{R}^p$. □

The above claim, being non-asymptotic and uniform in $\beta \in \mathbb{R}^p$, can be specialized to certain situations. In particular, it leads to an interesting conclusion under the asymptotic setting $p = p_n \to \infty$ and $\|\beta\|_0 \le s_n = o(p_n)$ as $n \to \infty$. Then the probability bound goes to 0 as $n \to \infty$, uniformly in $\beta \in \ell_0[s_n] \triangleq \{\beta : \|\beta\|_0 \le s_n\}$. Further, when $s_n = o(p_n)$, the constant C_0 can be chosen smaller, which makes the conclusion of the claim stronger.

Inference on β under the compatibility condition. The next several corollaries concern inference on β rather than on θ. Besides optimal prediction, it is of interest to infer on the parameter β itself. Because the dimension p may be (and generally is) larger than n, the correspondence between $\mathrm{X}\beta$ and β is not unique, and hence additional conditions are necessary even in the noiseless situation. As is commonly adopted in the literature (see, e.g., [42]), we will need to assume a condition lower bounding the norm of $\mathrm{X}\beta$ by a positive multiple of a norm on β for sparse vectors which is in turn a condition on the design matrix X.

There is yet another issue: recall that inference in the general framework is on θ and is based on the posterior $\hat{\pi}(\vartheta|Y)$ for $\theta = \mathrm{X}\beta$, not for β. In order to infer on β, we need to construct a prior Π on β that leads to an (empirical Bayes) posterior $\hat{\Pi}(b|Y)$ such that $\vartheta = \mathrm{X}b \sim \hat{\pi}(\vartheta|Y)$. This is not difficult: indeed, recall the construction (5.1) of the conditional prior $\pi_I(\vartheta|Y)$ on θ. Since in this case the unstructured $\theta = \mathrm{X}\beta$ and the conditional

5

prior $\pi_I(\vartheta|Y)$ was formally constructed as prior on $\theta^I = \mathrm{P}_I\theta = \mathrm{P}_I\mathrm{X}\beta = \mathrm{X}_I\beta$, we can derive the corresponding conditional prior $\Pi_I(b|Y)$ for $\beta = (\mathrm{X}_I^T\mathrm{X}_I)^{-1}\theta^I$ because $\theta^I = \mathrm{X}_I\beta$ is invertible with respect to β for any $I \in \mathcal{I}$. The corresponding conditional prior on β becomes

$$\beta|I \sim \Pi_I(b|Y) = \mathrm{N}\big((\mathrm{X}_I^T\mathrm{X}_I)^{-1}\mathrm{X}_I\mu, \kappa\sigma^2(\mathrm{X}_I^T\mathrm{X}_I)^{-1}\big) \otimes \delta_{0_{|I^c|}},$$

which means that subvector β_I with coordinates in I is normally distributed with the above parameters, and the remaining coordinates I^c of β are set to zero. From this point on, we can apply the (empirical) Bayesian approach in the same way as for θ. We thus construct the (empirical Bayes) posteriors on β: $\tilde{\Pi}_I(b|Y)$, $\Pi(b|Y)$, $\tilde{\Pi}(b|Y)$, $\check{\Pi}(b|Y)$; and the estimators $\tilde{\beta} = \sum_{I\in\mathcal{I}}\hat{\beta}_I\tilde{\pi}(I|Y)$ and $\check{\beta} = \hat{\beta}_{\hat{I}}$, where $\hat{I}$ is defined by (5.12) and $\hat{\beta}_I = (\mathrm{X}_I^T\mathrm{X}_I)^{-1}\mathrm{X}_I Y$ is just the ordinary least squares estimator of β based on the design matrix X_I of full column rank as $I \in \mathcal{I}$. Similarly, we can define $\hat{\Pi}(\beta|Y)$ as being either $\tilde{\Pi}(\beta|Y)$ or $\check{\Pi}(\beta|Y)$, and $\hat{\beta}$ as being either $\tilde{\beta}$ or $\check{\beta}$. The details of Bayesian construction for β can be found in [11]. For us what only matters is the fact that if $\beta \sim \hat{\Pi}(b|Y)$ then $\theta = \mathrm{X}\beta \sim \hat{\pi}(\vartheta|Y)$.

Introduce some additional notation. Recall that $\|\beta\|_0$ denotes the number of non-zero elements of β. Further let $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ be the ℓ_1-norm of β and $\|\mathrm{X}\|_{\max} = \max_{k=1,\dots,p}\|X_k\|$ (one can assume $\|\mathrm{X}\|_{\max} = \sqrt{n}$ without too much loss of generality). For $l \in \mathbb{N}$, let

$$\phi_1(l) = \inf\Big\{\frac{\sqrt{l}\|\mathrm{X}\beta\|}{\|\mathrm{X}\|_{\max}\|\beta\|_1} : \|\beta\|_0 \le l,\ \mathrm{supp}(\beta) \in \mathcal{I}\Big\}, \tag{5.59}$$

$$\phi_2(l) = \inf\Big\{\frac{\|\mathrm{X}\beta\|}{\|\mathrm{X}\|_{\max}\|\beta\|} : \|\beta\|_0 \le l,\ \mathrm{supp}(\beta) \in \mathcal{I}\Big\}. \tag{5.60}$$

Because $\|\beta\|_1 \le \sqrt{\|\beta\|_0}\|\beta\|$, it follows that $\phi_1(l) \ge \phi_2(l)$. Positivity of ϕ_1 at an argument l is called the compatibility condition, and is stronger if $\phi_1(l)$ is larger. If any of ϕ_1 or ϕ_2 is zero at its argument, then the corresponding result below becomes trivial but remains valid.

The following claims say basically that, under the compatibility condition, the (empirical Bayes) posterior $\hat{\Pi}(b|Y)$ on β contracts around the truth with the optimal rate.

Proposition 5.3. *Under the conditions of Corollary 5.1, for sufficiently large M_0' and any $M \ge 0$*

$$\begin{aligned}
&\mathbb{E}_\beta\hat{\Pi}\Big(\|b-\beta\|_1 \ge \tfrac{\sqrt{(C_0+1)\|\beta\|_0(M_0 r^2(\beta)+M\sigma^2)}}{\|\mathrm{X}\|_{\max}\phi_1((C_0+1)\|\beta\|_0)}\Big|Y\Big)\\
&\quad\le H_0 e^{-m_0 M} + C_v\exp\big\{-c_2(M_0'/2-c_3)\|\beta\|_0\log(\tfrac{ep}{\|\beta\|_0})\big\},\\
&\mathbb{E}_\beta\hat{\Pi}\Big(\|b-\beta\| \ge \tfrac{\sqrt{M_0 r^2(\beta)+M\sigma^2}}{\|\mathrm{X}\|_{\max}\phi_2((C_0+1)\|\beta\|_0)}\Big|Y\Big)\\
&\quad\le H_0 e^{-m_0 M} + C_v\exp\big\{-c_2(M_0'/2-c_3)\|\beta\|_0\log(\tfrac{ep}{\|\beta\|_0})\big\},
\end{aligned}$$

uniformly in $\beta \in \mathbb{R}^p$, where $C_0 = \max\{M_0', {M_0'}^2/e\}$.

Proof. By the definition of compatibility coefficient, on models I with $|I| \le C_0\|\beta\|_0$, the quantity $\|b-\beta\|_1$ is bounded by $\sqrt{(C_0+1)\|\beta\|_0}\|\mathrm{X}(b-\beta)\|/[\|\mathrm{X}\|_{\max}\phi_1((C_0+1)\|\beta\|_0)]$, since

the cardinality of $\mathrm{supp}(b-\beta)$ is at most $(C_0+1)\|\beta\|_0$. By Theorem 1, the $\mathbb{E}_\beta$-expectation of the posterior probability of $\|\mathrm{X}(b-\beta)\| = \|\vartheta-\theta\| > \sqrt{M_0 r^2(\theta)+M\sigma^2}$ is bounded by $H_0 e^{-m_0 M}$, while by Proposition 5.2, the event $\{I : |I| \ge C_0\|\beta\|_0\}$ has probability bounded by $C_v \exp\{-c_2(M_0'/2-c_3)\|\beta\|_0 \log(\frac{ep}{\|\beta\|_0})\}$. The first assertion follows, the proof of the second claim is similar. □

Notice that the above result implies the Corollary 5.4 in [42] and obtains optimal estimation rates for both ℓ_1 and ℓ_2 loss functions. Moreover, the dependence on the quantities $\phi_1(l)$ and $\phi_2(l)$ are optimal; cf. [73]. Next, we also obtain the optimal estimation result for both ℓ_1- and ℓ_2-norms.

Proposition 5.4. *Under the conditions of Corollary 5.1, for sufficiently large M_0' and any $M \ge 0$*

$$\begin{aligned}
&\mathbb{P}_\beta\Big(\|\hat\beta-\beta\|_1 \ge \tfrac{\sqrt{(C_0+1)\|\beta\|_0(M_1 r^2(\beta)+M\sigma^2)}}{\|\mathrm{X}\|_{\max}\phi_1((C_0+1)\|\beta\|_0)}\Big)\\
&\quad\le H_1 e^{-m_1 M} + C_v \exp\{-c_2(M_0'/2-c_3)\|\beta\|_0\log(\tfrac{ep}{\|\beta\|_0})\},\\
&\mathbb{P}_\beta\Big(\|\hat\beta-\beta\| \ge \tfrac{\sqrt{M_1 r^2(\beta)+M\sigma^2}}{\|\mathrm{X}\|_{\max}\phi_2((C_0+1)\|\beta\|_0)}\Big)\\
&\quad\le H_1 e^{-m_1 M} + C_v \exp\{-c_2(M_0'/2-c_3)\|\beta\|_0\log(\tfrac{ep}{\|\beta\|_0})\},
\end{aligned}$$

uniformly in $\beta\in\mathbb{R}^p$, where $C_0 = \max\{M_0', {M_0'}^2/e\}$.

Proof. Consider the case $\hat\theta = \check\theta = \mathrm{X}\check\beta$, where $\check\theta$ is defined by (5.13). Denote for brevity $\Delta = \frac{\sqrt{(C_0+1)\|\beta\|_0(M_1 r^2(\beta)+M\sigma^2)}}{\|\mathrm{X}\|_{\max}\phi_1((C_0+1)\|\beta\|_0)}$ and introduce the event $E_M = \{|\hat I| \le C_0\|\beta\|_0\}$, where $\hat I$ is defined by (5.12). By the definition of compatibility coefficient, in case $|\hat I| \le C_0\|\beta\|_0$, $\|\check\beta-\beta\|_1$ is bounded by $\sqrt{(C_0+1)\|\beta\|_0}\|\mathrm{X}(\check\beta-\beta)\|/(\|\mathrm{X}\|_{\max}\phi_1((C_0+1)\|\beta\|_0))$, since the cardinality of $\mathrm{supp}(\check\beta-\beta)$ is at most $(C_0+1)\|\beta\|_0$. By Theorem 1, $\|\mathrm{X}(\check\beta-\beta)\| > \sqrt{M_1 r^2(\beta)+M\sigma^2}$ has probability bounded by $H_1 e^{-m_1 M}$. Using this and Proposition 5.2, we have

$$\begin{aligned}
&\mathbb{P}_\beta(\|\check\beta-\beta\|_1 \ge \Delta) = \mathbb{P}_\beta(\|\check\beta-\beta\|_1 \ge \Delta, E_M) + \mathbb{P}_\beta(\|\check\beta-\beta\|_1 \ge \Delta, E_M^c)\\
&\quad\le \mathbb{P}_\beta(\|\mathrm{X}\check\beta-\mathrm{X}\beta\|^2 \ge M_1 r^2(\beta)+M\sigma^2) + \mathbb{P}_\beta(E_M^c) \le H_1 e^{-m_1 M} + \mathbb{P}_\beta(E_M^c)\\
&\quad\le H_1 e^{-m_1 M} + \mathbb{E}_\beta \check\pi(I : |I| > C_0\|\beta\|_0 | Y)\\
&\quad\le H_1 e^{-m_1 M} + C_v \exp\{-c_2(M_0'/2-c_3)\|\beta\|_0\log(\tfrac{ep}{\|\beta\|_0})\}.
\end{aligned}$$

The proof of the second claim for the case $\hat\theta = \check\theta = \mathrm{X}\check\beta$ and the proofs of the both claims for the case $\hat\theta = \tilde\theta = \mathrm{X}\tilde\beta$ are similar and therefore omitted. □

LINEAR REGRESSION WITH GROUP SPARSITY

Assume that the unknown regression vectors $\beta^1,\dots,\beta^m \in \mathbb{R}^p$ in (5.55) share the same support. Note that the model considered in Section 5.5.7 is a special case of linear regression with group sparsity with $m=1$. Local results for linear regression with group sparsity were derived in [61], and posterior contraction rate results in [42]. The group sparsity structure is modeled by the linear spaces

$$\mathbb{L}_I = \{\mathrm{vec}(\mathrm{X}_I^1 x_I^1,\dots,\mathrm{X}_I^m x_I^m) \in \mathbb{R}^{nm} : x_I^j \in \mathbb{R}^{|I|},\ j\in[m]\},\ \ I\in\mathcal{I},$$

where $\mathcal{I} = \mathcal{I}_1 \cup \{I_p\}$, with $I_p = [p]$, $\mathcal{I}_1 = \{I \subseteq [p] : m|I| + |I|\log(ep/|I|) \le r\}$, $r = \sum_{i=1}^m \operatorname{rank}(\mathrm{X}^i)$. Clearly, $|\mathcal{I}| \le 2^p$ and $d_I = \dim(\mathbb{L}_I) \le m|I|$ for $I \in \mathcal{I}_1$ and $d_{I_p} = r$. In this case, $\theta = \mathrm{X}\beta$ with $\beta = (\beta^1, \dots, \beta^m) \in \mathbb{R}^{mp}$, the structural slicing mapping is $s(I) = |I| \in \mathcal{S} \triangleq [p]_0$. Further, we have $|\mathcal{I}_{s(I)}| = \binom{p}{|I|}$ for $I \in \mathcal{I}_1$ and $|\mathcal{I}_{s(I_p)}| = 1$, hence $\log|\mathcal{I}_{s(I)}| = \log\binom{p}{|I|} \le |I|\log(ep/|I|)$ for $I \in \mathcal{I}_1$ and $\log|\mathcal{I}_{s(I_p)}| = 0$. Since $d_{s(I)} + \log|\mathcal{I}_{s(I)}| \le m|I| + |I|\log(ep/|I|)$ for $I \in \mathcal{I}_1$ and $d_{I_p} + \log|\mathcal{I}_{s(I_p)}| = r$, we take the majorant

$$\rho(s(I)) = \big(m|I| + |I|\log(ep/|I|)\big)1\{I \in \mathcal{I}_1\} + r1\{I = I_p\}, \quad I \in \mathcal{I}.$$

Notice the elbow effect in the majorant that emerges here for the same reason as in Section 5.5.7.

Conditions (A2) and (A3) are fulfilled in the same way as for the model in Section 5.5.7. As consequence of our general results, we obtain Corollary 5.1 for this case with the local rate $r^2(\beta) = \min_{I\in\mathcal{I}} r^2(I,\beta) = \min_{I\in\mathcal{I}} \{\|(\mathrm{I} - \mathrm{P}_I)\mathrm{X}\beta\|^2 + \sigma^2\rho(s(I))\}$.

Remark 5.21. *We can redefine the structural slicing mapping as $s(I) = \dim(\mathbb{L}_I)$, and the bound $\log|\mathcal{I}_s| = \log\binom{p}{s} \le s\log(ep/s)$ would still be valid. Notice further that we can slightly improve the above oracle rate by using the exact quantity $d_{s(I)} = d_I = \dim(\mathbb{L}_I)$ instead of its upper bound $m|I|$ in the expression for the complexity $\rho(s(I))$, which would make the oracle rate $r^2(\beta)$ also smaller.*

One can formulate the minimax results for the appropriate scales. For example, introduce the scale of classes

$$\ell_0^m[s] = \{\mathrm{vec}(\beta^1, \dots, \beta^m) \in \mathbb{R}^{pm} : I^*(\beta^i) = I^*(\beta^j), |I^*(\beta^i)| \le s \; \forall i, j \in [m]\},$$

where $I^*(\beta) = \{i \in [p] : \beta_i \ne 0\}$. The minimax rate over this class is established in [61] (under some conditions):

$$r^2(\ell_0^m[s]) \triangleq \inf_{\hat\beta} \sup_{\beta\in\ell_0^m[s]} \mathbb{E}_\beta\|\mathrm{X}\hat\beta - \mathrm{X}\beta\|^2 \asymp \sigma^2\big[ms + s\log(ep/s)\big].$$

Then we can easily show that the oracle rate implies this global rate, since

$$\begin{aligned} r^2(\beta) &\le r^2(I^*(\beta),\beta) \wedge r^2(I_p,\beta) \le \sigma^2\big[\big(m|I^*(\beta)| + |I^*(\beta)|\log(\tfrac{ep}{|I^*(\beta)|})\big) \wedge r\big] \\ &\le \sigma^2\big[ms + s\log(ep/s)\big] \asymp r^2(\ell_0^m[s]) \qquad \text{for all} \quad \beta \in \ell_0^m[s]. \end{aligned}$$

LINEAR REGRESSION WITH GROUP CLUSTERING

Assume now a clustering structure shared by m unknown regression vectors $\beta^1, \dots, \beta^m \in \mathbb{R}^p$. That is, there is some mapping $z : [m] \mapsto [k]$ such that $\beta^j = \beta^{z(j)}$, $j \in [m]$. Let the design matrix $\mathrm{X} = \mathrm{diag}\{\mathrm{X}^1, \dots, \mathrm{X}^m\}$ in (5.55) be such that $\mathrm{X}^1 = \dots = \mathrm{X}^m = \bar{\mathrm{X}}$, with $\det(\bar{\mathrm{X}}^T\bar{\mathrm{X}}) > 0$. Full column rankness of the $(n \times p)$-matrix $\bar{\mathrm{X}}$ implies $p \le n$. Each mapping $z \in [k]^{[m]}$ determines (uniquely) the pertinent partition $I = I(z) = (I_i, i \in [k])$ of the vectors $\beta^1, \dots, \beta^m$ into k groups $I_i = I_i(z) = z^{-1}(i) \subseteq [m]$, $i \in [k]$, such that $\cup_{i\in[k]} I_i = [m] = z^{-1}([k])$. Thus, the collection of all mappings $\mathcal{Z} = \mathcal{Z}(m) = \{z \in [k]^{[m]}, k \in [m]\}$ yields the collection of all clustering partitions of $[m]$: $\bar{\mathcal{I}} = \bar{\mathcal{I}}(m) = \{I(z), z \in [k]^{[m]}, k \in [m]\}$. Some local posterior contraction rate results for this model are claimed in [42], where this model is called by

multi-task learning. We will call this model rather by *linear regression with group clustering*. To the best of our knowledge, there are no adaptive minimax results on estimation and uncertainty quantification problems for this model.

In this model, the structures I are going to be certain partitions from $\bar{\mathcal{I}}$. Let $\bar{I} = (\{1\},\dots,\{m\})$ be the finest partition of $[m]$ into m one-point clusters and the structural slicing mapping $s(I)$ be the number of blocks in the partition I, so that $\mathcal{S} = [m]$. The group clustering structure is modeled by the following linear spaces

$$\mathbb{L}_I = \big\{\mathrm{vec}(\bar{\mathrm{X}}x^1,\dots,\bar{\mathrm{X}}x^m) \in \mathbb{R}^{nm} : x^j \in \mathbb{R}^p,\ j \in [m], \text{ such that } x^j = x^{j'}\ \forall\, j, j' \in I_i,\ I_i \in I,\ i \in [s(I)]\big\},$$

where $I \in \mathcal{I} \triangleq \mathcal{I}_1 \cup \{\bar{I}\}$ with $\mathcal{I}_1 = \{I \in \bar{\mathcal{I}} : ps(I) + m\log s(I) \le pm\}$. In this case, $\theta = \mathrm{X}\beta$, $d_I = \dim(\mathbb{L}_I) = ps(I)$ and $|\mathcal{I}_{s(I)}| = N(m, s(I))$ for $I \in \mathcal{I}_1$, where $N(m,s)$ is the number of ways to put m different objects into s different boxes so that each box contains at least one object. Then $\log|\mathcal{I}_{s(I)}| \le \log s^m(I) = m\log s(I)$ for $I \in \mathcal{I}_1$. Besides, we have $d_{\bar{I}} = \dim(\mathbb{L}_{\bar{I}}) = pm$ and $|\mathcal{I}_{s(\bar{I})}| = 1$. Since $d_{s(I)} + \log|\mathcal{I}_{s(I)}| \le ps(I) + m\log s(I)$ for $I \in \mathcal{I}_1$ and $d_{s(\bar{I})} + \log|\mathcal{I}_{s(\bar{I})}| = pm$, we take the majorant

$$\rho(s(I)) = \big(ps(I) + m\log s(I)\big)1\{I \in \mathcal{I}_1\} + pm1\{I = \bar{I}\}.$$

Remark 5.22. *As before, we have an elbow effect, again for the same reason. The idea of the elbow in the majorant should be clear now: there is no point (although possible) to model the structures $I \in \bar{\mathcal{I}}\backslash\mathcal{I}$, because all these structures are dominated by the structure $\bar{I} \in \mathcal{I}$. Indeed, for each $I \in \bar{\mathcal{I}}\backslash\mathcal{I}$, $r^2(I,\beta) = \|(\mathrm{I} - \mathrm{P}_I)\mathrm{X}\beta\|^2 + \sigma^2\rho(s(I)) = \|(\mathrm{I} - \mathrm{P}_I)\mathrm{X}\beta\|^2 + \sigma^2\big(ps(I) + m\log s(I)\big) \ge \sigma^2 pm = \|(\mathrm{I} - \mathrm{P}_{\bar{I}})\mathrm{X}\beta\|^2 + \sigma^2 pm = r^2(\bar{I},\beta)$, because $\mathrm{P}_{\bar{I}}\mathrm{X}\beta = \mathrm{X}\beta$.*

Condition (A2) is fulfilled, since, according to Remark 1.8, for any $\nu \ge 1$

$$\begin{aligned}\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} &\le \sum_{I\in\mathcal{I}_1} e^{-\nu\rho(s(I))} + e^{-\nu pm}\\ &\le \sum_{s\in[m]} e^{-\nu ps} + e^{-\nu pm} \le (e^{\nu p} - 1)^{-1} + 1 = C_\nu.\end{aligned}$$

Remark 5.23. *Notice that we could consider the full family of structures $\bar{\mathcal{I}}$ under some mild condition. Namely, we could allow redundancy by associating the same space $\mathbb{L}_{\bar{I}}$ to each $I \in \bar{\mathcal{I}}\backslash\mathcal{I}$. The majorant becomes $\bar{\rho}(s(I)) = \big(ps(I) + m\log s(I)\big)1\{I \in \mathcal{I}_1\} + pm1\{I \in \bar{\mathcal{I}}\backslash\mathcal{I}_1\}$, defined now for all $I \in \bar{\mathcal{I}}$. Then, if $p \gtrsim \log m$, Condition (A2) is fulfilled for sufficiently large ν:*

$$\begin{aligned}\sum_{I\in\bar{\mathcal{I}}} e^{-\nu\rho(s(I))} &\le \sum_{I\in\mathcal{I}_1} e^{-\nu\rho(s(I))} + \sum_{\epsilon\bar{\mathcal{I}}\backslash\mathcal{I}_1} e^{-\nu\rho(s(I))}\\ &\le \sum_{s\in[m]} e^{-\nu ps} + \sum_{s\in[m]} s^m e^{-\nu pm} \le (e^{\nu p} - 1)^{-1} + C = C_\nu.\end{aligned}$$

Thus, this structure redundancy $\bar{\mathcal{I}}\backslash\mathcal{I}_1$ does not affect the final local rate, only constant C_ν becomes slightly larger (and the condition $p \gtrsim \log m$ has to hold).

Condition (A3) is also fulfilled. Indeed, for any $I^0, I^1 \in \mathcal{I}$ define the partition refinement

$$I' = I'(I^0, I^1) = I^0 \vee I^1 = \big(I_i \cap J_j,\, I_i \in I^0, J_j \in I^1\big).$$

Clearly, $\mathbb{L}_{I^0} \cup \mathbb{L}_{I^1} \subseteq \mathbb{L}_{I'} \subseteq \mathbb{L}_{I^0} + \mathbb{L}_{I^1}$ and $\max\{s(I^0), s(I^1)\} \le s(I') \le s(I^0) + s(I^1)$, implying $\rho(s(I')) \le \rho(s(I^0)) + \rho(s(I^1))$, which entails Condition (A3).

As consequence of our general results, we obtain the local results of Corollary 5.1 for these model and structure with the local rate $r^2(\beta) = \min_{I\in\mathcal{I}}\{\|(\mathrm{I} - \mathrm{P}_I)\mathrm{X}\beta\|^2 + \sigma^2\rho(s(I))\}$.

In turn, by virtue of Corollary 5.2, the local results will imply global minimax adaptive results at once over all scales $\{\Theta_\gamma, \gamma \in \Gamma\}$ covered by the oracle rate $r^2(\beta)$ (i.e., for which (5.45) holds). For example, let $\Theta_{GC}(s) = \cup_{I\in\bar{\mathcal{I}}: s(I)\le s}\mathbb{L}_I$. To the best of our knowledge, there are no minimax results over $\Theta_{GC}(s)$. We conjecture that the minimax rate over $\Theta_{GC}(s)$ is

$$r^2(\Theta_{GC}(s)) \triangleq \inf_{\hat\beta} \sup_{\beta:\mathrm{X}\beta\in\Theta_{GC}(s)} \mathbb{E}_\beta\|\mathrm{X}\hat\beta - \mathrm{X}\beta\|^2 \asymp \sigma^2 \min\{ps + m\log s, pm\}.$$

It is not difficult to show that the local rate $r^2(\beta)$ covers this scale. Indeed, for each $\theta = \mathrm{X}\beta \in \Theta_{GS}(s)$ there exists $I_* = I_*(\theta) \in \bar{\mathcal{I}}$ such that $\theta = \mathrm{X}\beta \in \mathbb{L}_{I_*}$ and $s(I_*) \le s$. If $ps + m\log s \le pm$, then $I_* \in \mathcal{I}_1$. Hence, $r^2(\beta) \le r^2(I_*, \beta) = \sigma^2\rho(s(I_*)) = \sigma^2\big(ps(I_*) + m\log s(I_*)\big) \le \sigma^2(ps + m\log s)$, because $\mathrm{P}_{I_*}\mathrm{X}\beta = \mathrm{X}\beta$ and $s(I_*) \le s$. If $ps + m\log s > pm$, then $r^2(\beta) \le r^2(\bar I, \beta) = \sigma^2\rho(s(\bar I)) = \sigma^2 pm$, because $\mathrm{P}_{\bar I}\mathrm{X}\beta = \mathrm{X}\beta$.

Summarizing, $r^2(\beta) \le \sigma^2\min\{ps + m\log s, pm\}$. We thus established the relation (5.45) for this scale, and Corollary 5.2 follows with the minimax rate $r^2(\Theta_{GS}(s))$ defined above.

LINEAR REGRESSION WITH MIXTURE STRUCTURE

Consider the regression model (5.55) with $p \in [n]$ such that $\mathrm{X}^1 = \ldots = \mathrm{X}^m = \bar{\mathrm{X}}$, $\bar{\mathrm{X}} = (\bar X_{ij}) \in \{0,1\}^{n\times p}$, and $\sum_{j\in[p]}\bar X_{ij} = 1$ for all $i \in [n]$, i.e., each row of the matrix $\bar{\mathrm{X}}$ has $n-1$ zeros and only one entry equals to 1. Recently, some estimation results for this model were derived in [55]. To the best of our knowledge, there are no local results on posterior contraction rate and uncertainty quantification problems for mixture model.

In this case, $\theta = \mathrm{X}\beta$ and $\dim(\beta^j) = p \in [n]$, $j \in [m]$, is now not fixed but rather a varying ingredient of the structure. Another ingredient of the structure are the locations I_i of 1's in the ith p-dimensional row of the matrix $\bar{\mathrm{X}}$, $i \in [n]$. Putting these together, we encode the whole structure as $I = [p, (I_i, i \in [n])]$, where $I_i \in [p]$, $p \in [n]$. Thus, the full family of all structures is

$$\bar{\mathcal{I}} = \{[p, (I_i, i \in [n])] : I_i \in [p],\, p \in [n]\}.$$

Let $\mathrm{X}_I = \mathrm{diag}\{\bar{\mathrm{X}}_I, \ldots, \bar{\mathrm{X}}_I\}$, $\bar{\mathrm{X}}_I = (\bar X_{ij})$ be the $(n\times p(I))$-matrix corresponding to the structure $I \in \bar{\mathcal{I}}$, that is, $\bar X_{iI_i} = 1$ for $i \in [n]$ and all the other entries of this matrix are zeros. By $p(I)$ we denote the first ingredient of the structure I, the number of columns in the matrix $\bar{\mathrm{X}}_I$. The structural slicing mapping is $s(I) = r(I)$, where $r(I) = \mathrm{rank}(\bar{\mathrm{X}}_I)$, the number of linearly independent columns in the matrix $\bar{\mathrm{X}}_I$. So, $\mathcal{S} = [n]$ and notice that $r(I) \le p(I)$.

The structures in this model are modeled by the linear spaces

$$\mathbb{L}_I = \big\{\mathrm{vec}(\bar{\mathrm{X}}_I x^1, \ldots, \bar{\mathrm{X}}_I x^m) \in \mathbb{R}^{nm} : x^j \in \mathbb{R}^{p(I)}, j \in [m]\big\},$$

where $I \in \mathcal{I} \triangleq \mathcal{I}_1 \cup \{\bar{I}\}$ with $\mathcal{I}_1 = \{I \in \bar{\mathcal{I}} : mr(I) + n\log p(I) \le nm\}$ and $\bar{I} = \big[n, \lfloor n \rfloor\big]$ (so that $\bar{\mathrm{X}}_{\bar{I}} = \mathrm{I}$ is the n-dimensional identity matrix). In this case, $\theta = \mathrm{X}\beta$, $d_I = \dim(\mathbb{L}_I) = m\,\mathrm{rank}(\bar{\mathrm{X}}_I) = mr(I) \le mp(I)$ and $|\mathcal{I}_{s(I)}| \le p^n(I)$ for $I \in \mathcal{I}_1$, because $p^n(I)$ is the number of possibilities to choose locations of 1's in the n $p(I)$-dimensional rows of the design matrix $\bar{\mathrm{X}}_I$. Further, $d_{\bar{I}} = \dim(\mathbb{L}_{\bar{I}}) = nm$ (as $\bar{\mathrm{X}}_{\bar{I}} = \mathrm{I}$) and $|\mathcal{I}_{s(\bar{I})}| = 1$. Since $d_{s(I)} + \log|\mathcal{I}_{s(I)}| \le mr(I) + n\log p(I)$ for $I \in \mathcal{I}_1$ and $d_{s(\bar{I})} + \log|\mathcal{I}_{s(\bar{I})}| = nm$, we take the majorant

$$\rho(s(I)) = \big(mr(I) + n\log p(I)\big)1\{I \in \mathcal{I}_1\} + nm1\{I = \bar{I}\}. \tag{5.61}$$

The reason for considering the restricted family of structures $\mathcal{I}$ instead of the full family $\bar{\mathcal{I}}$ in this model is the same as for the model from Section 5.5.7 and is explained in Remark 5.22.

Condition (A2) is fulfilled since, according to Remark 1.8, for any $\nu \ge 1$

$$\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} \le \sum_{s\in\mathcal{S}} e^{-\nu d_s} \le \sum_{s\in[n]} e^{-\nu ms} + e^{-\nu nm} \le (e^{\nu m} - 1)^{-1} + 1 = C_\nu.$$

Condition (A3) can also be verified, which would ensure the coverage property (v) of Corollary 5.1 under EBR as well. However, there is no point in verifying Condition (A3), because for this linear regression model with mixture structure we have the same peculiar situation as for the biclustering model from Section 5.5.6: the size and coverage claims (vi)-(vii) for the confidence ball $B(\hat{\theta}, \tilde{R}_M)$ are stronger and more useful than the corresponding claims (iv)-(v) for the confidence ball $B(\hat{\theta}, \hat{R}_M)$. Let us demonstrate that the linear regression with mixture structure does not suffer from the deceptiveness phenomenon, modulo the so called highly structured parameters.

Indeed, as consequence of our general results, we obtain the local results (i)-(iv) and (vi)-(vii) of Corollary 5.1 for this case with the local rate $r^2(\beta) = \min_{I\in\mathcal{I}}\big\{\|(\mathrm{I} - \mathrm{P}_I)\mathrm{X}\beta\|^2 + \sigma^2\rho(s(I))\big\}$, with P_I as projection onto $\mathbb{L}_I$ defined above and the majorant $\rho(s(I))$ defined by (5.61). The coverage property (v) for the confidence ball $B(\hat{\theta}, \hat{R}_M)$ can be shown to hold also, but uniformly only under the EBR, whereas the coverage property (vii) for the confidence ball $B(\hat{\theta}, \tilde{R}_M)$ is uniform over the entire space $\Theta = \mathbb{R}^{n\times m}$. The size $\tilde{R}_M$ is of the oracle rate order (as the radius $\hat{R}_M$) uniformly in $\theta \in \Theta\backslash\tilde{\Theta} = \mathbb{R}^{n\times m}\backslash\tilde{\Theta}$, where $\tilde{\Theta}$ is defined by (5.28). Since in this model the total number of observations is $N = nm$, it is easy to see that $\tilde{\Theta} \subseteq \{\theta \in \mathbb{R}^{n\times m} : p(I_o(\theta)) = 1\}$ (i.e., $\bar{\mathrm{X}}_{I_o} = 1_n$, where 1_n is the n-dimensional column of 1's) where the oracle structure $I_o(\theta)$ is defined by (5.15). Clearly, the m-dimensional $\tilde{\Theta}$ is a "thin" subset of $\mathbb{R}^{n\times m}$ consisting of *highly structured parameters* θ whose oracle number of columns in the design matrix $\bar{\mathrm{X}}_{I_o}$ is $p(I_o(\theta)) = 1$. As we have already discussed at the end of Section 5.2.4, this means that, modulo these highly structured parameters, there is no deceptiveness phenomenon in this model.

Remark 5.24. *Notice that our local results for the linear regression model with mixture structure actually improve upon the results of [55] as we have $mr(I) \le mp(I)$ instead of $mp(I)$ (as in [55]) in the expression of the the local rate $r^2(\beta)$. This means that this oracle rate $r^2(\beta)$ defined above is smaller than the one from [55]. Notice that the below global minimax results over the considered class cannot be improved as the worst case of the both local rates is the same.*

Finally, by virtue of Corollary 5.2 the local results will imply global minimax adaptive results at once over all scales $\{\Theta_\gamma, \gamma \in \Gamma\}$ covered by the oracle rate $r^2(\beta)$ (i.e., for which (5.45) holds). Below we present one such scale, covered by the oracle rate $r^2(\beta)$.

Minimax results for mixture model. Define the class $\Theta_M(p) = \cup_{I\in\bar{\mathcal{I}}:p(I)\le p}\mathbb{L}_I$. As is shown in [55], the minimax rate over $\Theta_M(p)$ is

$$r^2(\Theta_M(p)) \triangleq \inf_{\hat\beta}\sup_{\beta:\mathrm{X}\beta\in\Theta_M(p)} \mathbb{E}_\beta\|\mathrm{X}\hat\beta-\mathrm{X}\beta\|^2 \asymp \sigma^2\min\{mp+n\log p, nm\}.$$

For each $\theta = \mathrm{X}\beta \in \Theta_M(p)$ there exists $I_* = I_*(\theta) \in \bar{\mathcal{I}}$ such that $\theta \in \mathbb{L}_{I_*}$ and $p(I_*) \le p$. If $mp + n\log p \le nm$, then $r(I_*) \le p(I_*) \le p$, hence $I_* \in \mathcal{I}_1$, so that $r^2(\beta) \le r^2(I_*, \beta) = \sigma^2\rho(s(I_*)) \le \sigma^2(mp + n\log p)$, because $\mathrm{P}_{I_*}\mathrm{X}\beta = \mathrm{X}\beta$. If $mp + n\log p > nm$, then $r^2(\beta) \le r^2(\bar I, \beta) = \sigma^2\rho(s(\bar I)) = \sigma^2 nm$, because $\mathrm{P}_{\bar I}\mathrm{X}\beta = \mathrm{X}\beta$.

Piecing these together, we obtain that $r^2(\beta) \le \sigma^2\min\{mp + n\log p, nm\}$ for all β such that $\theta = \mathrm{X}\beta \in \Theta_M(p)$. We thus established the relation (5.45) for this scale, and Corollary 5.2 follows with the minimax rate $r^2(\Theta_M(p))$ defined above.

5

5.5.8. AGGREGATION

Aggregation in nonparametric regression has been considered by [66], [27], [75], [84] and many others. We consider the regression model with a fixed design. The observation model is as follows:

$$Y_i = f(x_i) + \sigma\xi_i,\ i \in [n], \tag{5.62}$$

where $x_i \in \mathcal{X}$ are nonrandom, $\mathcal{X}$ is an arbitrary set, $f : \mathcal{X} \to \mathbb{R}$ is an unknown function, and $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. We use the notation $\|f\|^2 = \sum_{i\in[n]} f^2(x_i)$.

Assume we are given a collection of functions $\{f_1, ..., f_p\}$, called *dictionary*. For $\beta \in \mathbb{R}^p$, let $f_\beta = \sum_{j=1}^p \beta_j f_j$. By choosing a rich dictionary $\{f_1, ..., f_p\}$ and an appropriate $\beta \in \mathcal{B} \subseteq \mathbb{R}^p$, one can expect f_β to be close to f under some assumptions. For a certain choice of $\mathcal{B}$, the so called *aggregation problem* consists basically in determining the "best" $\hat\beta \in \mathcal{B}$ on the basis of the data Y such that $\left(\sum_{j=1}^p \hat\beta_j f_j(x_i), i \in [n]\right)$ well estimates the true $(f(x_i), i \in [n])$.

Introduce the sets $\mathcal{B}$ studied in the literature: the sets $\mathcal{B}_{(MS)}, \mathcal{B}_{(C)}, \mathcal{B}_{(L)}, \mathcal{B}_{(L_s)}, \mathcal{B}_{(C_s)}$ are defined as in [75]. Precisely, let $B_1(1) = \{\beta \in \mathbb{R}^p : \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \le 1\}$ and $B_0(s) = \ell_0[s] = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le s\}$ for $s \in [p]$. Next, define $\mathcal{B}_{(MS)} = B_0(1)$, $\mathcal{B}_{(C)}$ is a closed convex subset of $B_1(1)$, $\mathcal{B}_{(L)} = B_0(p) = \mathbb{R}^p$, $\mathcal{B}_{(L_s)} = B_0(s)$ and $\mathcal{B}_{(C_s)}$ as a closed convex subset of $B_0(s) \cap B_1(1)$. Thus $\mathcal{B} \in \{\mathcal{B}_{(MS)}, \mathcal{B}_{(C)}, \mathcal{B}_{(L)}, \mathcal{B}_{(L_s)}, \mathcal{B}_{(C_s)}\}$.

First recall the main estimation results from [75] (lower bounds are also established in that paper). The so called *exponential screening estimator* $f_{\tilde\beta^{\mathrm{ES}}}$ is proposed in [75]. Under the assumptions $\max_{j\in[p]}\|f_j\| \le \sqrt{n}$, $p \ge 2$, $n \ge 1$, $s \in [p]$, the following oracle estimation result is derived in [75]: for some constant $C > 0$,

$$\mathbb{E}_f\|f_{\tilde\beta^{\mathrm{ES}}} - f\|^2 \le \inf_{\beta\in\mathcal{B}}\|f_\beta - f\|^2 + C\sigma^2\psi_{n,p}(\mathcal{B}), \tag{5.63}$$

where $\psi_{n,p}(\mathcal{B}) = \min\{\psi^*_{n,p}(\mathcal{B}), r\}$ is the optimal rate of aggregation for the corresponding classes $\mathcal{B} \in \{\mathcal{B}_{(MS)}, \mathcal{B}_{(C)}, \mathcal{B}_{(L)}, \mathcal{B}_{(L_s)}, \mathcal{B}_{(C_s)}\}$, $r = \mathrm{rank}(\mathrm{X})$, $\psi^*_{n,p}(\mathcal{B})$ is defined as follows:

$$\psi^*_{n,p}(\mathcal{B}) = \begin{cases} \log p, & \mathcal{B} = \mathcal{B}_{(MS)}, \\ \sqrt{n \log\bigl(1 + \frac{ep\sigma}{\sqrt{n}}\bigr)}, & \mathcal{B} = \mathcal{B}_{(C)}, \\ r, & \mathcal{B} = \mathcal{B}_{(L)}, \\ s\log(1 + ep/s), & \mathcal{B} = \mathcal{B}_{(L_s)}, \\ \min\Bigl\{\sqrt{n \log\bigl(1 + \frac{ep\sigma}{\sqrt{n}}\bigr)}, s\log(1 + ep/s)\Bigr\}, & \mathcal{B} = \mathcal{B}_{(C_s)}. \end{cases}$$

An advantageous feature of the above results is its universality: the aggregation is attained over the five classes simultaneously. This result follows from Lemma 8.2 and Theorem 3.1 of [75]. Lemma 8.2 is fulfilled as soon as Theorem 3.1 holds and $\max_{j\in[p]} \|f_j\| \le \sqrt{n}$. The result of Theorem 3.1 from [75] in our notation reads as follows: for any $p, n \ge 1$

$$\mathbb{E}_f \|f_{\tilde{\beta}^{\mathrm{ES}}} - f\|^2 \le \min_{\beta\in\mathbb{R}^p} \Bigl\{\|f - f_\beta\|^2 + \sigma^2\bigl[r \wedge \bigl(9|I^*(\beta)|\log(1 + \tfrac{ep}{|I^*(\beta)|\vee 1})\bigr)\bigr]\Bigr\} + 8\sigma^2 \log 2. \tag{5.64}$$

A very useful fact concluded in [75] is that, under the condition $\max_{j\in[p]} \|f_j\| \le \sqrt{n}$, any estimator satisfying (5.64) (possibly with different constants in the right hand side) leads to the universal oracle inequality (5.63).

Let us demonstrate that we can derive the same type of estimation results as in [75], again as consequences of our general approach for particular choice of sparsity structures. In fact, we improve upon certain aspects and also provide the results on uncertainty quantification, again as consequence of our general framework results.

The aggregation problem considered here for the model (5.62) can be associated with the standard linear regression problem (5.55) (with $m = 1$) studied in Section 5.5.7. Indeed, let $\theta = (f(x_i), i \in [n])$ and notice that the vector $f_\beta = (f_\beta(x_i), i \in [n])$ can be represented as $\mathrm{X}\beta$, where $\beta \in \mathbb{R}^p$ is the unknown high-dimensional parameter and the design $(n \times p)$-matrix X has the entries $X_{ij} = f_j(x_i)$, $(i, j) \in [n] \times [p]$. In doing so, we arrive to the general setting $Y = \theta + \sigma\xi$, but now we take the same family of structures $\mathcal{I}$ and the corresponding family of linear spaces given by (5.56), as in Section 5.5.7. The conditions are already verified in Section 5.5.7. Then, according to Remark 5.18, the general results imply Corollary 5.1 with the oracle rate

$$r^2(\theta) = \min_{I\in\mathcal{I}} r^2(I, \theta) = \min_{I\in\mathcal{I}} \bigl\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho(s(I))\bigr\},$$

where the majorant $\rho(s(I))$ is defined by (5.57).

Recall the full family of structures $\tilde{\mathcal{I}} = \{J : J \subseteq [p]\}$. Since $\|\theta - \mathrm{P}_{I_r}\theta\|^2 = \min_{I\in\mathcal{I}} \|\theta - \mathrm{P}_I\theta\|^2 = \min_{I\in\tilde{\mathcal{I}}} \|\theta - \mathrm{P}_I\theta\|^2$, it is easy to see that

$$\begin{aligned} r^2(\theta) &= \bigl[\min_{I\in\mathcal{I}_1} r^2(I, \theta)\bigr] \wedge r^2(I_r, \theta) = \min_{I\in\tilde{\mathcal{I}}} \bigl\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2[\rho(s(I)) \wedge r]\bigr\} \\ &= \min_{\beta\in\mathbb{R}^p} \bigl\{\|f_\beta - f\|^2 + \sigma^2\bigl[r \wedge \bigl(2|I^*(\beta)|\log(\tfrac{ep}{|I^*(\beta)|})\bigr)\bigr]\bigr\}. \end{aligned} \tag{5.65}$$

5

In particular, property (ii) of Corollary 5.1 entails that for some $C_0, C_1 > 0$

$$\mathbb{E}_f \|f - \hat{\theta}\|^2 \le C_0 r^2(\theta) + C_1 \sigma^2,$$

where $r^2(\theta)$ is defined by (5.65), which is in fact property (5.64) for our estimator $\hat{\theta}$. As is mentioned above, [75] established that, under the additional assumption $\max_{j\in[p]} \|f_j\| \le \sqrt{n}$, this in turn leads to the universality property (5.63), now for the estimator $\hat{\theta}$: for some $C_0, C_2 > 0$,

$$\mathbb{E}_f \|f - \hat{\theta}\|^2 \le C_0 \inf_{\beta\in\mathcal{B}} \|f_\beta - f\|^2 + C_2 \sigma^2 \psi_{n,p}(\mathcal{B}).$$

We should mention that the constants in the universality property for our estimator $\hat{\theta}$ may be worse than those for the estimator $\tilde{f}_{\beta^{\mathrm{ES}}}$. On the other hand, notice that the claim (ii) of Corollary 5.1, being a uniform exponential inequality in probability, is itself finer and stronger versions of the corresponding oracle result in expectation (like (5.63)). Moreover, we additionally obtain claims (i) and (iv)-(vii) of Corollary 5.1 for the posterior concentration rate and uncertainty quantification. Global results over appropriate scales can also be derived as consequences of Corollary 5.2. Besides, we can drop the normality and independence assumptions and impose only Condition (A1) instead. One can readily formulate these results.

5

5.5.9. ISOTONIC, UNIMODAL AND CONVEX REGRESSIONS

Assume that the observations are $Y = (Y_i)_{i\in[n]}$ according to the model

$$Y_i = \theta_i + \sigma\xi_i, \ \ i \in [n],$$

where $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$ and $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ is the unknown high-dimensional parameter, possibly belonging to one of the three classes:

$$\mathcal{S}^{\uparrow} = \{\theta \in \mathbb{R}^n : \theta_i \le \theta_{i+1},\ i = 1, \dots, n-1\}, \ \ n \ge 2; \tag{5.66}$$

$$\mathcal{U}^m = \{\theta \in \mathbb{R}^n : \theta_1 \ge \dots \ge \theta_m \le \theta_{m+1} \le \dots \le \theta_n\}, \ m \in [n], \ n \ge 2; \tag{5.67}$$

$$\mathcal{C} = \{\theta \in \mathbb{R}^n : 2\theta_i \le \theta_{i+1} + \theta_{i-1},\ i = 2, \dots, n-1\}, \ \ n \ge 3. \tag{5.68}$$

The *isotonic*, *unimodal* and *convex* regression problems concern the classes $\mathcal{S}^{\uparrow}, \mathcal{U}^m$ and $\mathcal{C}$, respectively. Recently, oracle estimation results for these problems were derived by [35], [20] and [21]. To the best of our knowledge, there are no local results on posterior contraction rate and uncertainty quantification problems for these structures.

First, to model parameters from $\mathcal{S}^{\uparrow}$ and $\mathcal{U}^m$, introduce the linear spaces

$$\mathbb{L}_I = \{x \in \mathbb{R}^n : x_i = x_{i+1},\ i \notin I\}, \quad I \subseteq \mathcal{I} = [n-1],$$

where $d_I = \dim(\mathbb{L}_I) = |I| + 1$. The structural slicing mapping is $s(I) = |I|$, so that $\mathcal{S} = [n-1]$. Compute $|\mathcal{I}_{s(I)}| = \binom{n-1}{|I|}$, hence $\log|\mathcal{I}_{s(I)}| \le |I| \log(\frac{en}{|I|})$. Since $d_{s(I)} + |\mathcal{I}_{s(I)}| \le |I| + 1 + \log\binom{n-1}{|I|} \le 1 + 2|I|\log(\frac{en}{|I|})$, we take the majorant $\rho(s(I)) = 1 + 2|I|\log(\frac{en}{|I|})$.

Next, the parameters from $\mathcal{C}$ are modeled by the linear spaces

$$\mathbb{L}'_I = \{x \in \mathbb{R}^n : 2x_i = x_{i+1} + x_{i-1},\ i \notin I\}, \quad I \subseteq \mathcal{I}' = \{2, \dots, n-1\},$$

where $d_I = \dim(\mathbb{L}_I) \le (2|I| \vee 1) \wedge n \le 2|I|+1$. The structural slicing mapping in this case is $s'(I) = |I|$, so that $\mathcal{S}' = \{2, \dots n-1\}$. Compute $|\mathcal{I}_{s'(I)}| = \binom{n-2}{|I|}$, hence $\log|\mathcal{I}_{s'(I)}| \le |I|\log(\frac{en}{|I|})$. Since $d_{s'(I)} + |\mathcal{I}_{s'(I)}| \le 2|I|+1+|I|\log(\frac{en}{|I|}) \le 1+3|I|\log(\frac{en}{|I|})$, we take the majorant $\rho'(s'(I)) = 1+3|I|\log(\frac{en}{|I|})$.

We introduced two different families of structures with two corresponding (different) families of linear spaces, but the majorants in the both cases can be chosen the same (up to a multiplicative constant). Conditions (A1)-(A4) for the both cases are fulfilled in the same way as for the model considered in Section 5.5.3, we omit the argument and computations that are very much along the same lines as in Section 5.5.3. As consequence of our general results, we obtain the local results of Corollary 5.1 for the both cases with the local rate $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho(s(I))\}$ and $r'^2(\theta) = \min_{I\in\mathcal{I}'}\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho'(s'(I))\}$. In turn, by virtue of Corollary 5.2 the local results will imply global minimax adaptive results at once over all scales $\{\Theta_\beta, \beta\in\mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$ and $r'^2(\theta)$ (i.e., for which (5.45) holds). Below we present a couple of examples of such scales $\{\Theta_\beta, \beta\in\mathcal{B}\}$.

Minimax results for isotonic, unimodal and convex regressions up to a logarithmic factor. Following [20], for $\theta\in\mathbb{R}^n$, denote the number of relations $\theta_i \ne \theta_{i+1}$ for $i\in[n-1]$ by $k(\theta)-1$ (number of jumps of θ), and for $\theta\in\mathcal{C}$, the number of inequalities $2\theta_i \le \theta_{i+1}+\theta_{i-1}$ that are strict for $i = 2,\dots,n-1$ by $q(\theta)-1$. Let $\mathcal{U} = \cup_{m=1}^n \mathcal{U}^m$, where $\mathcal{U}^m$ is defined by (5.67). Define the classes of monotone and unimodal parameters with at most k jumps and the class of piecewise linear convex parameters with at most q linear pieces as follows: for $k, q \ge 1$,

$$\mathcal{S}_k^\uparrow = \{\theta\in\mathcal{S}^\uparrow : k(\theta)\le k\},\ \mathcal{U}_k = \{\theta\in\mathcal{U} : k(\theta)\le k\},\ \mathcal{C}_q = \{\theta\in\mathcal{C} : q(\theta)\le q\},$$

where $\mathcal{S}^\uparrow$ and $\mathcal{C}$ are defined by (5.66) and (5.68). Define $\Theta_k^\uparrow = \cup_{I\in\mathcal{I}:|I|+1\le k}\mathbb{L}_I$ and notice that for each $\theta\in\mathcal{S}_k^\uparrow$ (or $\theta\in\mathcal{U}_k$) there exists $I_*\in\mathcal{I}$ such that $\theta\in\mathbb{L}_{I_*}$ and $k(\theta) = |I_*|+1$, implying that $\mathcal{S}_k^\uparrow \subseteq \Theta_k^\uparrow$ (and $\mathcal{U}_k\subseteq\Theta_k^\uparrow$). Similarly, we define $\Theta_q = \cup_{I\in\mathcal{I}':|I|+1\le q}\mathbb{L}'_I$ and derive that $\mathcal{C}_q\subseteq\Theta_q$.

As is shown in [20], the minimax rates over $\mathcal{S}_k^\uparrow$ and $\mathcal{C}_q$, with $k, q\ge 1$, are $r^2(\mathcal{S}_k^\uparrow) \triangleq \inf_{\hat\theta}\sup_{\theta\in\mathcal{S}_k^\uparrow}\mathbb{E}_\theta\|\hat\theta-\theta\|^2 \asymp \sigma^2 k$ and $r^2(\mathcal{C}_q)\asymp\sigma^2 q$, respectively. Due to the fact that $\mathcal{S}^\uparrow\subseteq\mathcal{U}$, we also have $r^2(\mathcal{U}_k)\gtrsim\sigma^2 k$. Now, for each $\theta\in\mathcal{S}_k^\uparrow$ (or $\theta\in\mathcal{U}_k$) there exists $I_*\in\mathcal{I}$ such that $\theta\in\mathbb{L}_{I_*}$ (so that $\mathrm{P}_{I_*}\theta=\theta$) and $|I_*|+1 = k(\theta)\le k$. Hence, $r^2(\theta)\le r^2(I_*,\theta) = \sigma^2\big(1+2|I_*|\log(\frac{en}{|I_*|})\big) \lesssim \sigma^2 k\log(\frac{en}{k})$ for all $\theta\in\mathcal{S}_k^\uparrow$ and $\theta\in\mathcal{U}_k$. Similarly, we show that $r'^2(\theta)\lesssim \sigma^2 q\log(\frac{en}{q})$ for all $\theta\in\mathcal{C}_q$. We thus established the relation (5.45) for the classes $\mathcal{S}_k^\uparrow$, $\mathcal{U}_k$ and $\mathcal{C}_q$, which imply the minimax results (up to a logarithmic factor) of Corollary 5.2 for all these three classes.

Finally introduce the classes of (shape-restricted) monotone, unimodal and convex parameters θ with bounded total variation: $\mathcal{S}^\uparrow(V) = \{\theta\in\mathcal{S}^\uparrow : V(\theta)\le V\}$, $\mathcal{U}(V) = \{\theta\in\mathcal{U} : V(\theta)\le V\}$, $\mathcal{C}(V) = \{\theta\in\mathcal{C} : V(\theta)\le V\}$, where $V(\theta) = \max_{i,j}(\theta_i-\theta_j)$ (notice that $V(\theta) = \theta_n-\theta_1$ for $\theta\in\mathcal{S}^\uparrow$). It is known that the minimax rates over $\mathcal{S}^\uparrow(V)$, $\mathcal{U}(V)$ and $\mathcal{C}(V)$ are respectively $r^2(\mathcal{S}^\uparrow(V))\asymp\max\{n^{1/3}(\sigma^2V)^{2/3},\sigma^2\}$, $r^2(\mathcal{U}(V))\asymp\max\{n^{1/3}(\sigma^2V)^{2/3},\sigma^2\}$ and $r^2(\mathcal{C}(V))\gtrsim n^{1/5}(\sigma^4V)^{2/5}$ if $V\ge\sigma$. To derive the Corollary 5.2 for these classes, we

need the next proposition, where claim (i) is Lemma 2 from [20] and claim (ii) is Lemma 4.1 from [20]. We give these claims here (in our notation) for completeness, the proofs can be found in the mentioned references.

Proposition 5.5. *Let $k, q \in [n]$. Then the following properties hold.*

(i) For any $\theta \in \mathcal{S}^{\uparrow}$ (or $\theta \in \mathcal{U}$) there exists a $\theta^* = \theta^*(\theta) \in \Theta^{\uparrow}_k$ such that $\|\theta - \theta^*\|^2 \le C_1 \frac{nV^2(\theta)}{k^2}$ for some absolute constant $C_1 > 0$.

(ii) For any $\theta \in \mathcal{C}$ there exists a $\theta^* = \theta^*(\theta) \in \Theta_q$ such that $\|\theta - \theta^*\|^2 \le C_2 \frac{nV^2(\theta)}{q^4}$ for some absolute constant $C_2 > 0$.

According to Proposition 5.5, $\theta^* \in \Theta^{\uparrow}_k$, then there exists an $I_* \in \mathcal{I}$ such that $\theta^* \in \mathbb{L}_{I_*}$ and

$$\|\theta - \mathrm{P}_{I_*}\theta\|^2 \le \|\theta - \theta_*\|^2 \le C_1 \tfrac{nV^2(\theta)}{k^2}.$$

5

It follows therefore that for any $k \in [n]$ and any $\theta \in \mathcal{S}^{\uparrow}$ (or $\theta \in \mathcal{U}$) ,

$$r^2(\theta) \le r^2(I_*, \theta) \lesssim \tfrac{nV^2(\theta)}{k^2} + \sigma^2 k \log(\tfrac{en}{k}).$$

Let $\theta \in \mathcal{S}^{\uparrow}$ (or $\theta \in \mathcal{U}$) and let us take $k = k_* = \lfloor \big(\frac{nV^2(\theta)}{\sigma^2 \log(en)}\big)^{1/3} \rfloor + 1$. Now, if $k^* = 1$, we have $nV^2(\theta) \le \sigma^2 \log(en)$. If $k^* > 1$, then by definition of k^*, $\frac{nV^2(\theta)}{(k^*)^2} \le n^{1/3}(\sigma^2 V(\theta)\log(en))^{2/3}$. We conclude that if $V(\theta) \le V$, then for all $\theta \in \mathcal{S}^{\uparrow}(V)$ and $\theta \in \mathcal{U}(V)$,

$$r^2(\theta) \lesssim \max\{n^{1/3}(\sigma^2 V \log(en))^{2/3}, \sigma^2 \log(en)\}.$$

Thus, we established the relation (5.45) and Corollary 5.2 follows with the minimax rate $\max\{n^{1/3}(\sigma^2 V)^{2/3}, \sigma^2\}$ (up to a logarithmic factor) for the both classes $\mathcal{S}^{\uparrow}(V)$ and $\mathcal{U}(V)$ simultaneously.

Similarly, we establish that for all $\theta \in \mathcal{C}(V)$

$$r'^2(\theta) \lesssim \max\{n^{1/5}(\sigma^4 V)^{2/5}[\log(en)]^{4/5}, \sigma^2 \log(en)\}.$$

This means that we also established the relation (5.45) and hence also Corollary 5.2 with the minimax rate (up to a logarithmic factor) for the class $\mathcal{C}(V)$.

Below we present some further remarks and extensions for isotonic, unimodal and convex regressions.

Log factor as "price" for stronger results. It should be recognized that we attain the minimax rates for the classes $\mathcal{S}^{\uparrow}_k, \mathcal{U}_k, \mathcal{C}_q, \mathcal{S}^{\uparrow}(V), \mathcal{U}(V)$ and $\mathcal{C}(V)$ only up to a logarithmic factor. On the other hand, we obtain the optimal rates over the bigger scales $\{\Theta^{\uparrow}_k, k \in [n]\}$ and $\{\Theta_k, k \in [n]\}$. Moreover, as consequence of our general results we have also solved the uncertainty quantification problem and the problem of structure recovery (in a weak sense). Our constants in the estimation results may be worse than those from the above mentioned references, but on the other hand we do not require that the vector ξ is normal and its coordinates are independent, only mild Condition (A1) is to be fulfilled.

Universality of the results. Interestingly, that extra log factor in the local rate can also be seen as "price" for certain *universality* in the results. Indeed, recall that the results for the family of structures $\mathcal{I}$ with corresponding linear spaces $\mathbb{L}_I$, $I \in \mathcal{I}$, cover the scale $\{\Theta^{\uparrow}_k, k \in [n]\}$. This in turn implies the minimax results for the scales $\{\mathcal{S}^{\uparrow}_k, k \in [n]\}$ and $\{\mathcal{U}_k, k \in [n]\}$ (adaptively with respect to $k \in [n]$) and over the global shape-restricted classes $\mathcal{S}^{\uparrow}(V)$ and $\mathcal{U}(V)$ of monotone and unimodal parameters, *simultaneously* for all the mentioned scales. Thus, at the log factor price, one approach handles several structures at once.

Actually our approach allows to extend the universality property even further. Indeed, let us unite the two structures $\bar{\mathcal{I}} = \mathcal{I} \cup \mathcal{I}'$ and the corresponding families of the linear spaces $\{\mathbb{L}_I, I \in \bar{\mathcal{I}}\}$ and consider the resulting procedure. This procedure makes sense because the majorants for the both families are of the same order, so we only need to adjust a multiplicative constant in front of the majorant $\rho(s(I)) = 1 + 2|I|\log(\frac{en}{|I|})$ that will now handle the both families of structures. In doing so, we get the local result with the oracle rate over the both families at the price of a bigger multiple of the majorant. This means that the resulting procedure will mimic the oracle structure over the union of the two families, i.e., the resulting oracle rate will cover both scales $\{\Theta^{\uparrow}_k, k \in [n]\}$ and $\{\Theta_k, k \in [n]\}$ simultaneously. This in turn implies the minimax results for the scales $\{\mathcal{S}^{\uparrow}_k, k \in [n]\}$, $\{\mathcal{U}_k, k \in [n]\}$ and $\{\mathcal{C}_q, q \in [n]\}$ (adaptively with respect to $k, q \in [n]$) and over the global shape-restricted classes $\mathcal{S}^{\uparrow}(V)$, $\mathcal{U}(V)$ and $\mathcal{C}(V)$ of monotone, unimodal and convex parameters, *simultaneously* for all the mentioned scales.

No EBR-like condition for shape-restricted structures. The last important aspect to discuss for this model and the considered shape-restricted structures is one peculiar phenomenon recently discovered by some researchers in related settings: for certain shape-restricted classes, the uniform coverage and optimal size properties in the uncertainty quantification problem can be derived without imposing any EBR-like condition. For example, one can construct a confidence ball for monotone θ's with a high coverage and a radius of the optimal order $n^{1/3}(\sigma^2 V)^{2/3}$, uniformly over monotone $\theta \in \mathcal{S}^{\uparrow}(V)$ and without any EBR-like condition.

Let us show that we can also achieve this (up to a logarithmic factor) by using our approach. We consider only the family of structures $\mathcal{I}$ (and the corresponding family of linear spaces) for modeling monotone and unimodal θ's, similar argument can be given for the family of structures $\mathcal{I}'$. To ensure the EBR-condition, we simply restrict the family of structures $\mathcal{I}$ to the subfamily $\mathcal{I}_1 = \{I \in \mathcal{I} : |I| \geq C(n/\sigma^2)^{1/3}\}$ for some sufficiently large $C > 0$. Then the results go through in the same way as before with the difference that the oracle rate is now $r^2(\theta) = \min_{I \in \mathcal{I}_1} \{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho(s(I))\}$, with respect to the family $\mathcal{I}_1$, rather than $\mathcal{I}$. Since for any $I \in \mathcal{I}_1$, $|I| \geq C(n/\sigma^2)^{1/3}$. This and Proposition 5.5 imply that for any $I \in \mathcal{I}_1$ and any $\theta \in \mathcal{S}^{\uparrow}(V)$ (or $\theta \in \mathcal{U}(V)$) there exists a $\theta_* \in \Theta^{\uparrow}_I$ such that

$$\|\theta - \mathrm{P}_I\theta\|^2 \leq \|\theta - \theta_*\|^2 \leq C_1 \tfrac{nV^2}{|I|^2} \lesssim n^{1/3}\sigma^{4/3} \lesssim \sigma^2|I| \lesssim \sigma^2\rho(s(I)),$$

which ensures the EBR condition (5.24). At the same time, by taking $I_* \in \mathcal{I}_1$ such that $|I_*| = \lfloor C(n/\sigma^2)^{1/3} \rfloor + 1$, we obtain that uniformly over $\theta \in \mathcal{S}^{\uparrow}(V) \cup \mathcal{U}(V)$

$$r^2(\theta) \leq r^2(I_*, \theta) \leq C_1 \tfrac{nV^2}{|I_*|^2} + \sigma^2\rho(s(I_*)) \lesssim n^{1/3}\sigma^{4/3}\log(en),$$

which is the minimax rate (up to a logarithmic factor) over the both classes $\mathcal{S}^{\uparrow}(V)$ and $\mathcal{U}(V)$ simultaneously.

5.5.10. DICTIONARY LEARNING

Dictionary learning can be considered as a linear regression problem when the design matrix and (sparse) vector of regressors are both unknown. The data $Y = (Y_i)_{i\in[mn]}$ are observed according to the model:

$$Y = \bar{\mathrm{D}} r + \sigma\xi,$$

where $\xi = (\xi_i, i \in [mn]), \xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, $\bar{\mathrm{D}} = \mathrm{diag}\{\mathrm{D},\dots,\mathrm{D}\} \in \mathbb{R}^{mn\times mp}$ is a m block diagonal matrix with $p \in \mathbb{N}$, whose block $\mathrm{D} = (D_1,\dots,D_p) \in \mathbb{R}^{n\times p}$ is an unknown dictionary matrix, $p \le n$ without loss of generality, $\sigma > 0$ is the known noise intensity, $r = (r^1,\dots,r^m) \in \mathbb{R}^{mp}$ is a concatenation of unknown representations $r^1,\dots,r^m \in \mathbb{R}^p$ such that each entry r^j_i of each r^j comes from a (known) finite set of numbers: $r^j_i \in \mathcal{R}_K = \{\bar{r}_1,\dots,\bar{r}_K\}$ (for instance, $\mathcal{R}_3 = \{-1,0,1\}$), for some $\bar{r}_k \in \mathbb{R}$, $k \in [K]$. Recently, posterior contraction rate and oracle estimation results for this model were derived by [42] and [55], respectively. To the best of our knowledge, there are no local results on uncertainty quantification problem for dictionary learning.

In this model, we have $\theta = \bar{\mathrm{D}} r$. The structure I consists of two parts: m sparsity patterns $I^m \triangleq (I_1,\dots,I_m) \subseteq [p]^m$ (I_j determines which columns are taken in the j-th diagonal block D of $\bar{\mathrm{D}}$) and m sparse versions of representation vectors $R_{I^m} \triangleq (r^1_{I_1},\dots,r^m_{I_m})$ according to the sparsity patterns I^m, where $r^j_{I_j} = (r^j_i, i \in I_j)$ with $r^j_i \in \mathcal{R}_K$, $i \in I_j$, $j \in [m]$. We encode the structure I as $I = (I^m, R_{I^m})$, and the whole family of structures is

$$\bar{\mathcal{I}} = \{(I^m, R_{I^m}) : r^j_i \in \mathcal{R}_K,\ i \in I_j,\ j \in [m];\ I_k \subseteq [p],\ k \in [m]\}.$$

The structural slicing mapping is defined as $s(I) = (|I_k|, k \in [m]) \in \mathcal{S} \triangleq [p]^m_0$. Further, introduce the subfamily $\mathcal{I}_1$ of $\bar{\mathcal{I}}$:

$$\mathcal{I}_1 = \{I \in \bar{\mathcal{I}} : np + l_K(I) \le nm\},$$

where the quantity $l_K(I)$ is defined as

$$l_K(I) \triangleq \sum_{j\in[m]} |I_j| \log\left(\tfrac{ep}{|I_j|}\right) + (\log K) \sum_{j\in[m]} |I_j|. \tag{5.69}$$

This quantity has the meaning of the log of the cardinality of the structural layer $\mathcal{I}_{s(I)}$ and its motivation to appear here will become clear later.

The structures in this model are modeled by the linear spaces

$$\mathbb{L}_I = \left\{\mathrm{vec}(\mathrm{D}_{I_1} r^1_{I_1},\dots,\mathrm{D}_{I_m} r^m_{I_m}) \in \mathbb{R}^{nm} : \mathrm{D}_{I_k} \in \mathbb{R}^{n\times|I_k|},\ k \in [m]\right\},$$

where $I \in \mathcal{I} \triangleq \mathcal{I}_1 \cup \{\bar{I}\}$ and $\bar{I}$ is one special structure (the finest possible) such that $s(\bar{I}) = (p,\dots,p)$ (m-dimensional vector of p's) and the associated linear space is $\mathbb{L}_{\bar{I}} =$

$\{\mathrm{vec}(x^1,\dots,x^m)\in\mathbb{R}^{nm}: x^j\in\mathbb{R}^n,\ j\in[m]\}$. If some $I_j=\varnothing$, then the corresponding column $\mathrm{D}_{I_j}r^j_{I_j}$ is the zero column.

In this case, $\theta=\bar{\mathrm{D}}r\in\mathbb{R}^{n\times m}$ (recall that whenever appropriate we treat θ as vector: $\theta\in\mathbb{R}^{nm}$), $d_I=\dim(\mathbb{L}_I)=n|\cup_{k\in[m]}I_k|\le np$ for $I\in\mathcal{I}_1$ and $d_{\bar I}=\dim(\mathbb{L}_{\bar I})=nm$. The layer $\mathcal{I}_{s(I)}$ consists of all the structures I which have the same $s(I)=(|I_k|,k\in[m])$. Clearly, $|\mathcal{I}_{s(\bar I)}|=1$ because there is only one structure $\bar I$ in the layer $\mathcal{I}_{s(\bar I)}$. To count the number of structures in $\mathcal{I}_{s(I)}$ for $I\in\mathcal{I}_1$, notice that there are $\prod_{j\in[m]}\binom{p}{|I_j|}$ possible choices of the sparsity patterns I^m and there are $K^{\sum_{k\in[m]}|I_k|}$ possible choices of sparse representation vectors R_{I^m}, yielding the cardinality $|\mathcal{I}_{s(I)}|=\prod_{j\in[m]}\binom{p}{|I_j|}\times K^{\sum_{k\in[m]}|I_k|}$. Hence,

$$\log|\mathcal{I}_{s(I)}|\le\sum_{j\in[m]}|I_j|\log(\tfrac{ep}{|I_j|})+(\log K)\sum_{j\in[m]}|I_j|=l_K(I)\quad\text{for } I\in\mathcal{I}_1,$$

where $l_K(I)$ is introduced by (5.69). The last relation explains the origin of the quantity $l_K(I)$. Since $d_{s(I)}+\log|\mathcal{I}_{s(I)}|\le np+l_K(I)$ for $I\in\mathcal{I}_1$ and $d_{s(\bar I)}+\log|\mathcal{I}_{s(\bar I)}|=nm$, we take the majorant

$$\rho(s(I))=\big(np+l_K(I)\big)1\{I\in\mathcal{I}_1\}+nm1\{I=\bar I\}. \tag{5.70}$$

As for some previous models and structures, we have an elbow effect expressed basically by the quantity $l_K(I)$ in the majorant and there is no need to consider the structures $I\in\bar{\mathcal{I}}\backslash\mathcal{I}$, because these structures are dominated by the structure $\bar I$, the same reasoning applies as in Remark 5.22.

Conditions (A1) and (A4) hold with $d_{s(I)}=\dim(\mathbb{L}_I)$ in view of Remarks 1.6 and 5.10. Denote $\mathcal{S}_1=\{s(I):I\in\mathcal{I}_1\}$. Condition (A2) is fulfilled, since, according to Remark 1.8, for a sufficiently large $\nu>1$

$$\begin{aligned}\sum_{I\in\mathcal{I}}e^{-\nu\rho(s(I))}&\le\sum_{I\in\mathcal{I}_1}e^{-\nu\rho(s(I))}+e^{-\nu nm}\le e^{-\nu np}\sum_{s\in\mathcal{S}_1}e^{-(\nu-1)l_K(I)}+e^{-\nu nm}\\&\le e^{-\nu np}\sum_{|I_1|=0}^{p}\dots\sum_{|I_m|=0}^{p}e^{-(\nu-1)l_K(I)}+e^{-\nu nm}\\&\le e^{-\nu np}\Big(\sum_{l=0}^{p}e^{-(\nu-1)l}\Big)^m+1\le\tfrac{e^{-\nu np}}{(1-e^{1-\nu})^m}+1\le C_\nu,\end{aligned}$$

under the assumption that $m\lesssim np$.

Remark 5.25. *Notice the emerging condition $m\lesssim np$. This is not completely surprising: m should not be too big in order not to have too many structures in the layers. Alternatively, instead of imposing this condition, we can make the majorant slightly bigger by setting $np\vee m$ instead of just np in (5.70). Yet another fix would be to remove those structures I from $\mathcal{I}_1$ for which $\sum_{j\in[m]}|I_j|<m$. One can show that in this case the above sum will be uniformly bounded.*

As for the previous model (linear regression with mixture structure), there is no point in verifying Condition (A3) because the size and coverage claims (vi)-(vii) for the confidence ball $B(\hat\theta,\tilde R_M)$ are stronger and more useful for this model and this structure than

the corresponding claims (iv)-(v) for the confidence ball $B(\hat{\theta}, \hat{R}_M)$. Let us demonstrate that this model in essence does not suffer from the deceptiveness phenomenon, modulo the so called highly structured parameters.

Indeed, as consequence of our general results, we obtain the local results (i)-(iv) and (vi)-(vii) of Corollary 5.1 for this case with the local rate $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|(\mathrm{I}-\mathrm{P}_I)\theta\|^2 + \sigma^2\rho(s(I))\}$, where $\theta = \bar{\mathrm{D}}r$, with majorant $\rho(s(I))$ defined above and P_I, the projection onto $\mathbb{L}_I$ defined above. The coverage property (v) for the confidence ball $B(\hat{\theta}, \hat{R}_M)$ can be shown to hold also, but uniformly only under the EBR, whereas the coverage property (vii) for the confidence ball $B(\hat{\theta}, \tilde{R}_M)$ is uniform over the entire space $\theta = \bar{\mathrm{D}}r \in \Theta = \mathbb{R}^{n\times m}$. The size $\tilde{R}_M$ is of the oracle rate order (as the radius $\hat{R}_M$) uniformly in $\theta \in \Theta\backslash\tilde{\Theta} = \mathbb{R}^{n\times m}\backslash\tilde{\Theta}$, where $\tilde{\Theta}$ is defined by (5.28). In this model the total number of observations is $N = nm$ and $\tilde{\Theta} = \{\theta \in \mathbb{R}^{n\times m} : \sigma^{-2}\|(\mathrm{I}-\mathrm{P}_{I_o(\theta)})\theta\|^2 + np + l_K(I_o(\theta)) \lesssim \sqrt{N} = \sqrt{nm}\}$, where $I_o(\theta)$ is the oracle structure defined by (5.15). Clearly, $\tilde{\Theta}$ is a "thin" subset of $\mathbb{R}^{n\times m}$ consisting of *highly structured parameters* θ, in this case *ultra-sparse* parameters as their oracle structure must be very sparse: $l_K(I_o(\theta)) \le C\sqrt{nm} - np$. Actually, $\tilde{\Theta} = \varnothing$ if $m \lesssim p^2 n$ which is a very mild assumption on the dimensions n, p, m only. To summarize, under the assumption $m \lesssim p^2 n$, in the dictionary learning model there is no deceptiveness issue at all.

Remark 5.26. *Notice that we actually established stronger local results: the local rate* $r^2(\theta) = \min_{I\in\mathcal{I}}\{\|(\mathrm{I}-\mathrm{P}_I)\theta\|^2 + \sigma^2\bar{\rho}(s(I))\}$ *is with a smaller majorant* $\bar{\rho}(s(I)) = \min\{n|\cup_{i\in[m]} I_i| + \sum_{i\in[m]}|I_i|\log(\frac{ep}{|I_i|}) + (\log K)\sum_{i\in[m]}|I_i|, nm\}$, *under the assumption* $m \lesssim np$. *If we want to avoid the assumption* $m \lesssim np$, *then we should put* $\big(n|\cup_{i\in[m]} I_i|\big) \vee m$ *instead of* $n|\cup_{i\in[m]} I_i|$ *in the expression of the majorant* $\bar{\rho}(s(I))$.

Finally, by virtue of Corollary 5.2 the local results will imply global minimax adaptive results at once over all scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$ (i.e., for which (5.45) holds). Below we present one example of scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$.

Minimax results for sparse dictionary learning. Define the sparsity class for the dictionary learning model: for $\bar{s} \in [p]_0$, $\Theta_{SDL}(\bar{s}) = \cup\{\mathbb{L}_I : I \in \bar{\mathcal{I}}, |I_i| \le \bar{s}, i \in [m]\}$. As is shown in [55], the minimax rate over $\Theta_{SDL}(\bar{s})$ is

$$r^2(\Theta_{SDL}(\bar{s})) \asymp \sigma^2 \min\{np + m\bar{s}\log(\tfrac{ep}{\bar{s}}), nm\}.$$

For each $\theta = \bar{\mathrm{D}}r \in \Theta_{SDL}(\bar{s})$ there exists $I_* \in \bar{\mathcal{I}}$ such that $\theta \in \mathbb{L}_{I_*}$, hence $\mathrm{P}_{I_*}\theta = \theta$ and $r^2(I_*(\theta),\theta) = \sigma^2\rho(s(I_*))$. Further, since $|I_{*i}(\theta)| \le \bar{s}$, $i \in [m]$, we have $l_K(I_*(\theta)) = \sum_{i\in[m]}|I_{*i}(\theta)|\log(\frac{ep}{|I_{*i}(\theta)|}) + (\log K)\sum_{i\in[m]}|I_{*i}(\theta)| \le (1+\log K)m\bar{s}\log(\frac{ep}{\bar{s}})$. Therefore, if $np + (1+\log K)m\bar{s}\log(\frac{ep}{\bar{s}}) \le nm$, then $np + l_K(I_*(\theta)) \le np + (1+\log K)m\bar{s}\log(\frac{ep}{\bar{s}}) \le nm$, hence $I_*(\theta) \in \mathcal{I}_1$ and $r^2(\theta) \le r^2(I_*(\theta),\theta) = \sigma^2\rho(s(I_*(\theta))) = \sigma^2[np + l_K(I_*(\theta))] \le \sigma^2[np + (1+\log K)m\bar{s}\log(\frac{ep}{\bar{s}})]$ in this case. Besides, recall that $\mathrm{P}_{\bar{I}}\theta = \theta$, so that $r^2(\theta) \le r^2(\bar{I},\theta) = \sigma^2\rho(s(\bar{I})) = \sigma^2 nm$. Piecing these together, we obtain that

$$r^2(\theta) \lesssim \sigma^2 \min\{np + m\bar{s}\log(\tfrac{ep}{\bar{s}}), nm\} \asymp r^2(\Theta_{SDL}(\bar{s})) \text{ for all } \theta \in \Theta_{SDL}(\bar{s}).$$

We thus established the relation (5.45) for this scale, and Corollary 5.2 follows with the minimax rate $r^2(\Theta_{SDL}(\bar{s}))$ defined above.

5.5.11. MEAN MATRIX WITH SUBMATRIX SPARSITY

Suppose we observe a matrix $Y = (Y_{ij}) \in \mathbb{R}^{n_1 \times n_2}$:

$$Y_{ij} = \theta_{ij} + \sigma\xi_{ij}, \quad i \in [n_1], \quad j \in [n_2],$$

where $\sigma > 0$ is the known noise intensity, $\xi_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, $\theta = (\theta_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ is an unknown high-dimensional parameter of interest with at most k_1 nonzero rows and k_2 nonzero columns, which are not necessarily consecutive. To the best of our knowledge, there are no local results on estimation, posterior contraction rate and uncertainty quantification problems for this model.

The submatrix sparsity structure is modeled by the linear subspaces

$$\mathbb{L}_I = \big\{\mathrm{vec}(x) \in \mathbb{R}^{n_1 n_2} : x_{ij} = 0 \ \ \forall (i,j) \in \big((I_1^c \times [n_2]) \cup ([n_1] \times I_2^c)\big)\big\},$$

where $I = (I_1, I_2) \in \mathcal{I} = \{(I_1', I_2') : I_1' \subseteq [n_1], I_2' \subseteq [n_2]\}$ and $d_I = \dim(\mathbb{L}_I) = |I_1||I_2|$. The structural slicing mapping is $s(I) = (|I_1|, |I_2|)$, so that $\mathcal{S} = ([n_1]_0, [n_2]_0)$. Compute $|\mathcal{I}_{s(I)}| = \prod_{i=1}^2 \binom{n_i}{|I_i|}$, hence $\log|\mathcal{I}_{s(I)}| = \log\binom{n_1}{|I_1|} + \log\binom{n_2}{|I_2|} \le \sum_{i=1}^2 |I_i| \log(\frac{en_i}{|I_i|})$. Since $d_{s(I)} = d_I = |I_1||I_2|$ and $d_{s(I)} + \log|\mathcal{I}_{s(I)}| \le d_I + \sum_{i=1}^2 |I_i| \log(\frac{en_i}{|I_i|})$, we take the majorant $\rho(s(I)) = |I_1||I_2| + |I_1| \log(\frac{en_1}{|I_1|}) + |I_2| \log(\frac{en_2}{|I_2|})$.

Conditions (A1) and (A4) hold with $d_{s(I)} = \dim(\mathbb{L}_I)$ in view of Remarks 1.6 and 5.10. Condition (A2) is fulfilled, since, according to Remark 1.8, for any $v > 1$

$$\begin{aligned}\sum_{I \in \mathcal{I}} e^{-v\rho(s(I))} &\le \sum_{|I_1|=0}^{n_1} \big(\tfrac{en_1}{|I_1|}\big)^{-(v-1)|I_1|} \sum_{|I_2|=0}^{n_2} \big(\tfrac{en_2}{|I_2|}\big)^{-(v-1)|I_2|} \\ &\le \tfrac{1}{(1-e^{1-v})^2} = C_v.\end{aligned}$$

For any $I^0, I^1 \in \mathcal{I}$ define $I' = I'(I^0, I^1) = (I_1^0 \cup I_1^1, I_2^0 \cup I_2^1)$. Then $(\mathbb{L}_{I^0} \cup \mathbb{L}_{I^1}) \subseteq \mathbb{L}_{I'}$ and $\rho(s(I')) \le \rho(s(I^0)) + \rho(s(I^1))$, which entails Condition (A3).

As consequence of our general results, we obtain the local results of Corollary 5.1 for this case with the local rate $r^2(\theta) = \min_{I \in \mathcal{I}} \big\{\|\theta - \mathrm{P}_I\theta\|^2 + \sigma^2\rho(s(I))\big\}$. In turn, by virtue of Corollary 5.2 the local results will imply global minimax adaptive results at once over all scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$ (i.e., for which (5.45) holds). Below we present the example of scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\theta)$.

Minimax results for $\mathcal{F}(k_1, k_2, n_1, n_2)$. Let $\mathcal{F}(k_1, k_2, n_1, n_2)$ be the collection of matrices $\theta = (\theta_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ with at most k_1 nonzero rows and k_2 nonzero columns, which are not necessarily consecutive. Classes $\mathcal{F}(k_1, k_2, n_1, n_2)$ were introduced in [62]. In our notation, $\mathcal{F}(k_1, k_2, n_1, n_2) = \cup_{I \in \mathcal{I}: |I_1| \le k_1, |I_2| \le k_2} \mathbb{L}_I$. As is shown in [62], the minimax rate over $\mathcal{F}(k_1, k_2, n_1, n_2)$ is

$$r^2(\mathcal{F}(k_1, k_2, n_1, n_2)) \asymp \sigma^2\big(k_1k_2 + k_1\log(\tfrac{en_1}{k_1}) + k_2\log(\tfrac{en_2}{k_2})\big).$$

On the other hand, for each $\theta \in \mathcal{F}(k_1, k_2, n_1, n_2)$ there exists $I_* \in \mathcal{I}$ such that $\theta \in \mathbb{L}_{I_*}$ and $|I_{*1}| \le k_1$ and $|I_{*2}| \le k_2$. Hence, $\mathrm{P}_{I_*}\theta = \theta$ and

$$\begin{aligned}r^2(\theta) \le r^2(I_*, \theta) = \sigma^2\rho(s(I_*)) &= \sigma^2\big(|I_{*1}||I_{*2}| + |I_{*1}|\log(\tfrac{en_1}{|I_{*1}|}) + |I_{*2}|\log(\tfrac{en_2}{|I_{*2}|})\big) \\ &\le \sigma^2\big(k_1k_2 + k_1\log(\tfrac{en_1}{k_1}) + k_2\log(\tfrac{en_2}{k_2})\big) \asymp r^2(\mathcal{F}(k_1, k_2, n_1, n_2)).\end{aligned}$$

We thus established the relation (5.45) for this scale, and Corollary 5.2 follows with the minimax rate $r^2(\mathcal{F}(k_1,k_2,n_1,n_2))$ defined above.

5.5.12. COVARIANCE MATRIX WITH BANDING OR SPARSITY STRUCTURE

Suppose we observe n iid p-dimensional vectors $X_1,\dots,X_n$, $X_i=(X_i^1,\dots,X_i^p)^T$, $i\in[n]$, with $\mathbb{E}X_i=0$, $\mathbb{E}(X_i^j)^4\le C_X$, $(i,j)\in[p]\times[p]$, and unknown covariance matrix $\mathbb{E}(X_iX_i^T)=\Sigma$, $i\in[n]$. Without loss of generality, we set $C_X=1$ for the rest of this section. Let $\mathcal{C}\subseteq\mathbb{R}^{p\times p}$ denote the set of all p-dimensional covariance matrices. Assume that for some (known and independent of p) $\varepsilon_0>0$,

$$\Sigma\in\mathcal{C}_{\varepsilon_0}=\{\mathrm{M}\in\mathcal{C}:\varepsilon_0\le\lambda_{\min}(\mathrm{M})\le\lambda_{\max}(\mathrm{M})\le\varepsilon_0^{-1}\}.$$

Here, $\lambda_{\max}(\mathrm{M})$ and $\lambda_{\min}(\mathrm{M})$ are the maximum and minimum eigenvalues of M. We assume that $X_i\overset{\text{ind}}{\sim}\mathrm{N}(0_n,\Sigma)$, where 0_n is the n-dimensional vector of zeros. The normality assumption is not important to us, this only plays a role in that we can use certain auxiliary result below (Proposition 5.6) which is available only for the normal case.

We are interested in recovering the covariance matrix $\Sigma=\{\Sigma_{ij}\}_{1\le i,j\le p}$ which is assumed to have the *banding* or *sparsity* structure, to be specified later. The maximum likelihood estimator of Σ is $\tilde\Sigma=\frac{1}{n}\sum_{l=1}^n(X_l-\bar X)(X_l-\bar X)^T=\frac{1}{n}\sum_{l=1}^nX_lX_l^T-\bar X\bar X^T$, where $\bar X=\frac{1}{n}\sum_{l=1}^nX_l$. Since $\bar X\bar X^T$ is a higher order term (see Remark 1 in [29]), we shall ignore this term and focus on the dominating term $\frac{1}{n}\sum_{l=1}^nX_lX_l^T$ for estimating the covariance matrix Σ.

Let $\mathrm{Y}=(Y_{ij})_{i,j\in[p]}=\frac{1}{n}\sum_{l=1}^nX_lX_l^T$, $Y=\mathrm{vec}\big[(Y_{ij})\big]=(Y_{11},Y_{12}\dots,Y_{pp})^T$. We obtained the following model:

$$Y_{ij}=\Sigma_{ij}+\sigma_n\xi_{ij},\quad i,j\in[p],\tag{5.71}$$

where $\sigma_n\xi_{ij}=Y_{ij}-\mathbb{E}Y_{ij}=Y_{ij}-\Sigma_{ij}$, so that $\mathbb{E}\xi_{ij}=0$ and

$$\sigma_n^2\mathrm{Var}(\xi_{ij})=\tfrac{1}{n}\mathrm{Var}(X_1^iX_1^j)=\tfrac{1}{n}\mathbb{E}(X_1^iX_1^j)^2\le\tfrac{1}{n}[\mathbb{E}(X_1^i)^4]^{\frac12}[\mathbb{E}(X_1^j)^4]^{\frac12}\le\tfrac{1}{n}.$$

The parameter σ_n will be chosen later, for now it is any sequence $\sigma_n\in[0,1]$. We thus have a particular case of general framework model (1.15), where the parameter of interest is now denoted by Σ instead of θ. Recall that we work with the usual norm of vectorized version of the parameter $\Sigma=(\Sigma_{ij})_{i,j\in[p]}$, that is, if Σ is seen as matrix, then $\|\Sigma\|$ means its Frobenius norm. We denote the probability measure of Y from the model (5.71) by $\mathbb{P}_\Sigma$, and the corresponding expectation $\mathbb{E}_\Sigma$.

BANDING STRUCTURE

Assume that the covariance matrix $\Sigma=(\Sigma_{ij})_{i,j\in[p]}$ has a *banding* structure, i.e., $\Sigma_{ij}=0$ for all $i,j\in[p]$ such that $|i-j|>I$ for some $I\in[p]_0$. To model this structure, define the linear spaces

$$\mathbb{L}_I=\big\{\mathrm{vec}(x)\in\mathbb{R}^{p^2}:x_{ij}=x_{ji}\ \forall i,j\in[p];x_{ij}=0\text{ if}|i-j|>I\big\},\ I\in\mathcal{I}=[p]_0.$$

Then $\|\Sigma - P_I\Sigma\|^2 = \sum_{|i-j|>I}\Sigma_{ij}^2$, $d_I = \dim(\mathbb{L}_I) = p + \sum_{l=1}^{I}(p-l) = p + I(p-(I+1)/2)$, the structural slicing mapping $s(I) = I$, $\mathcal{S} = [p-1]_0$, $\log|\mathcal{I}_s| = 0$, $d_{s(I)} = p + I(p-(I+1)/2)$ leading to the majorant $\rho(s(I)) = d_I = p + I(p-(I+1)/2)$.

Condition (A2) is fulfilled, since $\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} \le \sum_{s\in\mathcal{S}} e^{-\nu s} \le \frac{e^{\nu}}{e^{\nu}-1} = C_\nu$ for any $\nu > 0$. However, in order to derive at least the local estimation and posterior contraction results, we also need Condition (A1). This condition is now not easy to check since the errors ξ_{ij}'s are dependent in the model (5.71). We apply the following strategy (in the same spirit as in Section 5.5.5): introduce certain event and establish that the probability of this event is exponentially small (in n); next, under this event establish Condition (A1); finally, combine these two facts to derive the local estimation and posterior contraction results.

The following proposition (formulated in our notation) is Lemma 12 from Appendix of supplementary material [57] and is given here for completeness, its proof can be found in [57].

Proposition 5.6. *Let* $\nu_{ij} = \max\{(\Sigma_{ii}\Sigma_{jj})^{1/2} - \Sigma_{ij}, (\Sigma_{ii}\Sigma_{jj})^{1/2} + \Sigma_{ij}\}$, $i, j \in [p]$. *Then for any* $t \in [0, \nu_{ij}/2)$

$$\mathbb{P}(\sigma_n|\xi_{ij}| \ge t) \le 4\exp\Big\{-\frac{3nt^2}{16\nu_{ij}^2}\Big\}.$$

The relation $\mathbb{P}(\max_{i,j\in[p]}|\xi_{ij}| \ge t) \le \sum_{i,j\in[p]}\mathbb{P}(|\xi_{ij}| \ge t)$ and Proposition 5.6 imply that, for the event $E = \{\max_{i,j\in[p]}|\xi_{ij}| \le t_0\}$ with $t_0 = \frac{\min_{i,j\in[p]}\nu_{ij}}{\sqrt{5}}$,

$$\mathbb{P}(E^c) = \mathbb{P}\Big(\max_{i,j\in[p]}|\xi_{ij}| \ge t_0\Big) \le H'\exp\{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2\log p\}, \tag{5.72}$$

where ν_{ij} is defined in Proposition 5.6, $H' = 4$, $0 < \bar{c}_1 = \frac{3\varepsilon_0^4}{320} \le \frac{3}{80}\frac{\min_{i,j\in[p]}\nu_{ij}^2}{\max_{i,j\in[p]}\nu_{ij}^2}$ (because $\varepsilon_0 \le \min_{i,j\in[p]}\nu_{ij} \le \max_{i,j\in[p]}\nu_{ij} \le 2\varepsilon_0^{-1}$) and $\bar{c}_2 = 2$. For (5.72) to be useful, we need to assume the asymptotic relation $\log p = o(n\sigma_n^2)$ as $n\to\infty$.

By the assumptions on Σ, we have that $\min_{i,j\in[p]}\nu_{ij} \le 2\varepsilon_0^{-1}$, so that $t_0^2 = \min_{i,j\in[p]}\nu_{ij}^2/5 \le \frac{4}{5\varepsilon_0^2}$. Using this and (5.72), we ensure Condition (A1) under the event $E = \{\max_{i,j\in[p]}|\xi_{ij}| \le t_0\}$ with $\alpha = 1\wedge(5\varepsilon_0^2)/4$. Exactly,

$$\begin{aligned}\mathbb{E}\exp\big\{\alpha\|P_I\xi\|^2\big\}1\{E\} &= \mathbb{E}\exp\Big\{\alpha\sum_{|i-j|>I}\xi_{ij}^2\Big\}1\{\max_{i,j\in[p]}|\xi_{ij}| < t_0\}\\ &\le \exp\big\{\alpha t_0^2(p + I(p-(I+1)/2))\big\} \le \exp\{d_{s(I)}\}.\end{aligned} \tag{5.73}$$

Condition (A3) holds as well. Indeed, for any $I_0, I_1 \in \mathcal{I}$ take $I' = I_0 \vee I_1$ and verify that $(\mathbb{L}_{I_0}\cup\mathbb{L}_{I_1}) \subseteq \mathbb{L}_{I'} = \mathbb{L}_{I_0} + \mathbb{L}_{I_1}$ and $\rho(s(I')) \le \rho(s(I_0)) + \rho(s(I_1))$.

The oracle rate is in this case $r^2(\Sigma) = \min_{I\in\mathcal{I}} r^2(I,\Sigma)$, where

$$r^2(I,\Sigma) = \|\Sigma - P_I\Sigma\|^2 + \sigma_n^2\rho(s(I)) = \sum_{|i-j|>I}\Sigma_{ij}^2 + \sigma_n^2\big(p + I(p-(I+1)/2)\big),$$

and the EBR-set $\Theta_{\mathrm{eb}} = \Theta_{\mathrm{eb}}(t)$ is given by (5.24), but now in terms of the bias and variance parts of the oracle rate $r^2(\Sigma)$.

We have thus verified the conditional version of Condition (A1) (under the event E) and Conditions (A2) and (A3) for the model (5.71) with the banding structure. This means that we can derive results on estimation, posterior contraction and uncertainty quantification for this model. These are the counterparts of claims (i)-(v) of Corollary 5.1 summarized by Theorem 5.6 below. To the best of our knowledge, there are no local results on estimation, posterior contraction rate and uncertainty quantification problems for this model.

A couple of conventions concerning notation in Theorem 5.6: as compared to the general framework notation, in the model (5.71), the parameter of interest is denoted by Σ instead of θ and the corresponding estimator becomes $\hat{\Sigma}$ instead of $\hat{\theta}$; in the posteriors for Σ we use the variable Σ' to distinguish it from the "true" $\Sigma \in \mathcal{C}_{\varepsilon_0}$. We keep the same notation for all other quantities involved as in the general framework (like $\hat{r}$, $\hat{R}_M$, $B(\hat{\Sigma}, \hat{R}_M)$), with the understanding that these are specialized for the model (5.71) with the banding structure and the oracle rate $r^2(\Sigma)$.

Theorem 5.6. *Let the constants $M_0, M_1, M_3, H_0, H_1, H_2, H_3, m_0, m_1, m_2, m_3, c_2, c_3, C_v, H', \bar{c}_1, \bar{c}_2$ be defined in Theorems 5.1-5.3 and (5.72). Then for any $M \geq 0$,*

$$\sup_{\Sigma \in \mathcal{C}_{\varepsilon_0}} \mathbb{E}_\Sigma \hat{\pi}\big(\|\Sigma' - \Sigma\|^2 \geq M_0 r^2(\Sigma) + M\sigma_n^2 | Y\big) \leq H' e^{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2 \log p} + H_0 e^{-m_0 M},$$

$$\sup_{\Sigma \in \mathcal{C}_{\varepsilon_0}} \mathbb{P}_\Sigma\big(\|\hat{\Sigma} - \Sigma\|^2 \geq M_1 r^2(\Sigma) + M\sigma_n^2\big) \leq H' e^{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2 \log p} + H_1 e^{-m_1 M},$$

$$\sup_{\Sigma \in \mathcal{C}_{\varepsilon_0}} \mathbb{E}_\Sigma \hat{\pi}\big(I : r^2(I, \Sigma) \geq c_3 r^2(\Sigma) + M\sigma_n^2 | Y\big) \leq H' e^{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2 \log p} + C_v e^{-c_2 M},$$

$$\sup_{\Sigma \in \mathcal{C}_{\varepsilon_0}} \mathbb{P}_\Sigma\big(\hat{r}^2 \geq M_3 r^2(\Sigma) + (M+1)\sigma_n^2\big) \leq H' e^{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2 \log p} + H_3 e^{-m_3 M},$$

$$\sup_{\Sigma \in \mathcal{C}_{\varepsilon_0} \cap \Theta_{\mathrm{eb}}(t)} \mathbb{P}_\Sigma\big(\Sigma \notin B(\hat{\Sigma}, \hat{R}_M)\big) \leq H' e^{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2 \log p} + H_2 e^{-m_2 M}.$$

Let us outline the idea of the proof (which is omitted) of the first claim of the above theorem; the same reasoning applies to the remaining claims. The expectation of the empirical Bayes posterior probability $\mathbb{E}_\Sigma \Pi = \mathbb{E}_\Sigma \hat{\pi}\big(\|\Sigma' - \Sigma\|^2 \geq M_0 r^2(\Sigma) + M\sigma_n^2 | Y\big)$ is bounded by the sum of two terms $\mathbb{E}_\Sigma \Pi \leq \mathbb{P}_\Sigma(E^c) + \mathbb{E}_\Sigma \Pi 1_E$. The first term is evaluated by using (5.50) (obtaining the bound $H' e^{-\bar{c}_1 n\sigma_n^2 + \bar{c}_2 \log p}$); the second term is evaluated exactly in the same way as in the proof Theorem 5.1, because Condition (A1) is fulfilled under the event E according to (5.73). Counterparts of assertions (ii) and (iii) of Theorem 5.2 can also be formulated and proved in the same way.

As to the choice of σ_n^2, this quantity is in the oracle rate, so that we would want it to be as small as possible. On the other hand, we want the claims of the theorem to be non-void, which is ensured only if $\sigma_n^2 n \geq C \log p$, or $\sigma_n^2 \geq \frac{C \log p}{n}$, for sufficiently large $C > 0$. In the sequel we take therefore $\sigma_n^2 = \frac{C \log p}{n}$. An extra log factor thus appeared which will also enter the minimax rates in the global results. We conjecture that one can get rid of that factor by using more accurate concentration inequalities when establishing Condition (A1).

As usually, the local results of Theorem 5.6 will imply global minimax adaptive results at once over all scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\Sigma)$ (i.e., for which (5.45)

holds). Below we present the example of scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the oracle rate $r^2(\Sigma)$.

Minimax results for the scale $\{\mathcal{G}_\beta, \beta > 0\}$. For $\beta, L, \varepsilon_0 > 0$, define

$$\mathcal{G}_\beta = \mathcal{G}_\beta(L, \varepsilon_0^{-1}) = \big\{\Sigma \in \mathcal{C}_{\varepsilon_0} : |\Sigma_{ij}| \le L|i-j|^{-(\beta+1)} \text{ for } i \ne j\big\}.$$

The rate $r^2(\mathcal{G}_\beta) = \min\big\{pn^{-\frac{2\beta+1}{2(\beta+1)}}, p^2n^{-1}\big\}$ is minimax over the class $\mathcal{G}_\beta$ under the Frobenius norm; see [29]. If $(\frac{n}{\log p})^{\frac{1}{2(\beta+1)}} \le p$, taking $I_* = \lfloor (n/\log p)^{\frac{1}{2(\beta+1)}} \rfloor$ and recalling $\sigma_n^2 = \frac{C\log p}{n}$, we derive that, uniformly in $\Sigma \in \mathcal{G}_\beta$,

$$\begin{aligned} r^2(\Sigma) \le r^2(I_*, \Sigma) &= \sum_{|i-j|>I_*} \Sigma_{ij}^2 + \sigma_n^2\big(p + I_*(p - (I_*+1)/2)\big) \\ &\lesssim pI_*^{-(2\beta+1)} + \frac{pI_*\log p}{n} \lesssim p\big(\frac{n}{\log p}\big)^{-\frac{2\beta+1}{2(\beta+1)}}. \end{aligned}$$

If $(\frac{n}{\log p})^{\frac{1}{2(\beta+1)}} > p$, we take $I_* = p$ to derive $\sup_{\Sigma\in\mathcal{G}_\beta} r^2(\Sigma) \lesssim p^2(\frac{n}{\log p})^{-1}$.

To summarize, we established that

$$\sup_{\Sigma\in\mathcal{G}_\beta} r^2(\Sigma) \lesssim \min\big\{p(\frac{n}{\log p})^{-\frac{2\beta+1}{2(\beta+1)}}, p^2\big(\frac{n}{\log p}\big)^{-1}\big\} = \tilde{r}^2(\mathcal{G}_\beta),$$

where $\tilde{r}^2(\mathcal{G}_\beta)$ is the minimax rate (up to a logarithmic factor) for the class $\mathcal{G}_\beta$. Then the last relation and Theorem 5.6 imply the global minimax results for the scale $\{\mathcal{G}_\beta, \beta > 0\}$. These results will look as the ones from Theorem 5.6 with the difference that the class $\mathcal{G}_\beta$ stands instead $\mathcal{C}_{\varepsilon_0}$ and the rate $\tilde{r}^2(\mathcal{G}_\beta)$ stands instead of $r^2(\Sigma)$. For the results to be most useful, we take $\sigma_n^2 = \frac{C\log p}{n}$ with sufficiently large $C > 0$ and $M = M_n \to \infty$ as $n \to \infty$ such that $M_n\sigma_n^2 \asymp \tilde{r}^2(\mathcal{G}_\beta)$.

Remark 5.27. *The obtained local and global results on uncertainty quantification for the covariance matrix with a banding structure are new to the best of our knowledge. Notice however that we derived only the uncertainty quantification results based on the EBR condition, whereas counterparts of claims (vi)-(vii) of Corollary 5.1 could not be established because we were unable to verify Condition (A4). The point is that the set $\tilde{\Theta}$ of highly structured parameters defined by (5.28) is empty in this case: $r^2(\Sigma) \ge \sigma_n^2\rho(s(I)) \gtrsim \sigma_n^2 p = \sigma_n^2 N^{1/2}$ (as $N = p^2$). This means that the uncertainty quantification claims based on Condition (A4) would be more valuable for this model because they are free of the deceptiveness phenomenon. Indeed, if we would have established Condition (A4), then the confidence ball $B(\hat{\Sigma}, \tilde{R}_M)$ would have been of asymptotically full coverage and of the optimal oracle size, uniformly over $\mathcal{C}_{\varepsilon_0}$ (because $\tilde{\Theta}$ turns out to be empty in this case). It is an open problem to verify Condition (A4) for the model (5.71), the main issue is to find an appropriate statistics $V(Y')$ for which the second relation of Condition (A4) is fulfilled.*

SPARSITY STRUCTURE

Here we briefly discuss the case of sparsity structure for the model (5.71).

Denote by Σ_{-i} the i-th column of Σ with Σ_{ii} removed. Let $p \geq 2$. For any $i \in [p]$ the vector $\Sigma_{-i} \in \mathbb{R}^{p-1}$ is assumed to be *sparse* so that $\Sigma_{ki} = 0$, $k \notin I_i$ (or $\Sigma_{ik} = 0$), where $I_i \subseteq [p]\backslash\{i\}$. To model this sparsity structure, introduce the linear spaces

$$\mathbb{L}_I = \{\mathrm{vec}(x) \in \mathbb{R}^{p^2} : x_{ki} = 0,\ k \notin I_i,\ i \in [p]\},$$

where the structure is $I = (I_1, \dots, I_p) \in \mathcal{I} \triangleq \{(J_1, \dots, J_p) : J_i \subseteq [p]\backslash\{i\}, i \in [p]\}$. Then $\|\Sigma - \mathrm{P}_I\Sigma\|^2 = \sum_{i\in[p]}\sum_{j\notin I_i}\Sigma_{ji}^2$, $\dim(\mathbb{L}_I) = p + \sum_{i=1}^p |I_i|$. The structural slicing mapping is $s(I) = (|I_1|, \dots, |I_p|) \in \otimes_{i\in[p]}[p-1]_0 = \mathcal{S}$, $\log|\mathcal{I}_{s(I)}| = \sum_{i\in[p]}\log\binom{p-1}{|I_i|} \leq \sum_{i\in[p]}|I_i|\log\left(\frac{e(p-1)}{|I_i|}\right)$, $d_{s(I)} = p + \sum_{i\in[p]}|I_i|$. Thus, we take the majorant $\rho(s(I)) = p + \sum_{i\in[p]}|I_i|\log\left(\frac{e(p-1)}{|I_i|}\right)$.

Condition (A2) is fulfilled, since, according to Remark 1.8,

$$\sum_{I\in\mathcal{I}} e^{-\nu\rho(s(I))} \leq e^{-\nu p}\sum_{s_1=0}^{p-1} e^{-(\nu-1)s_1} \dots \sum_{s_p=0}^{p-1} e^{-(\nu-1)s_p} \leq \frac{e^{-\nu p}}{(1-e^{1-\nu})^p} = C_\nu,$$

for sufficiently large $\nu > 1$.

Next, along the same lines as in Section 5.5.12, we can verify the conditional version of Condition (A1) (under event E) and Conditions (A2) and (A3) for the model (5.71) with the sparsity structure. This means that we can derive results on estimation, posterior contraction and uncertainty quantification for the model (5.71), now with the sparsity structure. We can readily formulate a theorem containing the local results for this structure: it will take the form of Theorem 5.6 with the oracle rate $r^2(\Sigma) = \min_{I\in\mathcal{I}}\{\|\Sigma - \mathrm{P}_I\Sigma\|^2 + \sigma_n^2\rho(s(I))\}$, where $\sigma_n^2 = \frac{C\log p}{n}$ with sufficiently large $C > 0$. To the best of our knowledge, there are no local results on estimation, posterior contraction rate and uncertainty quantification problems for this model. Also for this sparsity structure we have the same issue (described in Remark 5.27) with Condition (A4) as for the banding structure.

Finally, consider one scale covered by the oracle rate for the sparsity structure.

Weak ℓ_q-balls. Recall the weak ℓ_q ball of radius c in $\mathbb{R}^m$ containing elements with fast decaying ordered magnitudes of components,

$$B_q^m(c) = \{\zeta \in \mathbb{R}^m : |\zeta|_{(k)}^q \leq ck^{-1},\ k \in [m]\},$$

where $|\zeta|_{(k)}$ denotes the kth largest element in magnitude of the vector ζ. For $0 \leq q < 1$, define the class $\mathcal{G}_q(c_{n,p})$ of covariance matrices by

$$\mathcal{G}_q(c_{n,p}) = \{\Sigma \in \mathcal{C}_{\varepsilon_0} : \Sigma_{-j} \in B_q^{p-1}(c_{n,p}),\ j \in [p]\},$$

that is, each column Σ_{-j} of $\Sigma \in \mathcal{G}_q(c_{n,p})$ must be in a weak ℓ_q ball, $j \in [p]$. The minimax estimation rate over this class is $r^2(\mathcal{G}_q(c_{n,p})) = pc_{n,p}\left(\frac{\log p}{n}\right)^{1-q/2} + \frac{p}{n}$; see [30]. Recall $\sigma_n^2 = \frac{C\log p}{n}$ and take $I^* = I^*(\Sigma) = (I_1^*, \dots, I_p^*)$ such that $|I_i^*| = p^* \triangleq \lfloor c_{n,p}\left(\frac{\log p}{n}\right)^{-q/2}\rfloor$, $i \in [p]$, to

derive

$$\begin{aligned}\sup_{\Sigma\in\mathcal{G}_q(c_{n,p})} r^2(\Sigma) &\le \sup_{\Sigma\in\mathcal{G}_q(c_{n,p})} r^2(I^*,\Sigma)\\ &\le \sup_{\Sigma\in\mathcal{G}_q(c_{n,p})} \sum_{i\in[p]}\sum_{j\notin I_i^*}\Sigma_{ji}^2+\sigma_n^2\big[p+pp^*\log\big(\tfrac{e(p-1)}{p^*}\big)\big]\\ &\lesssim pc_{n,p}^{2/q}\sum_{j>p^*} j^{-2/q}+\sigma_n^2 p\, c_{n,p}n^{q/2}(\log p)^{1-q/2}+\sigma_n^2 p\\ &\lesssim \big(pc_{n,p}\big(\tfrac{\log p}{n}\big)^{1-q/2}+\tfrac{p}{n}\big)\log p.\end{aligned}$$

This relation and the local results imply the global minimax results (up to the logarithmic factor $\log p$) on estimation, posterior contraction and uncertainty quantification for the model (79) for the scale $\mathcal{G}_q(c_{n,p})$.

REFERENCES

[1] ABRAMOVICH, F., BENJAMINI, Y. DONOHO, D.L. and JOHNSTONE, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* 34, 584–653.

[2] ABRAMOVICH, F., GRINSHTEIN, V. and PENSKY, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* 35, 2261–2286.

[3] AIROLDI, E.M. and CHAN, S.H. (2014). A Consistent Histogram Estimator for Exchangeable Graph Models. *J. Mach. Learn. Res.* 32, 208–216.

[4] AIROLDI, E.M., COSTA, T.B. and CHAN, S.H. (2013). Stochastic block model approximation of a graphon: Theory and consistent estimation. *Adv. Neural Inf. Process. Syst.* 26, 692–700.

[5] BABENKO, A. and BELITSER, E. (2010). Oracle projection convergence rate of posterior. *Math. Meth. Statist.* 19, 219–245.

[6] BALAZC, G., GYÖRGY, A. and SZEPESVARI, S. (2015). Near-optimal max-affine estimators for convex regression. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, PLMR 38, 56–64.

[7] BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* 32, 528–551.

[8] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* 113, 301-413.

[9] BELITSER, E. (2017). On coverage and local radial rates of credible sets. *Ann. Statist.* 45, 1124–1151.

[10] BELITSER, E. and GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite dimensional normal distribution. *Ann. Statist.* 31, 536–559.

[11] BELITSER, E. and GHOSAL, S. (2018). Empirical Bayes oracle uncertainty quantification for regression. Preprint.

[12] BELITSER, E. and LEVIT, B. (1995). On minimax filtering over ellipsoids. *Math. Meth. Statist.* 3, 259–273.

[13] BELITSER, E. and NURUSHEV, N. (2018). Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences. Under revision for *Bernoulli* (see also ArXiv:1511.01803).

[14] BELITSER, E. and NURUSHEV, N. (2017). Local posterior concentration rate for multilevel sparse sequences. *Bayesian Statistics in Action*, Springer Proc. Math. Stat. 194, 51–66.

[15] BELITSER, E. and NURUSHEV, N. (2017). Discussion on article by van der Pas, Szabo, and van der Vaart. *Bayesian Analysis* 12, 1267–1269.

[16] BELITSER, E. and NURUSHEV, N. (2018). Local inference by penalization method for biclustering model. *Math. Methods Statist.* 27, 163–183.

[17] BELITSER, E. and NURUSHEV, N. (2018). Robust inference for general projection structures by empirical Bayes and penalization methods. *Submitted.*

[18] BELITSER, E., NURUSHEV, N. and SERRA, P. (2018). Robust estimation of sparse signal with unknown sparsity cluster value. *Submitted.*

[19] BELKIN, M., MATVEEVA, I. and NIYOGI, P. (2004). Regularization and semi-supervised learning on large graphs. *COLT*, Springer 3120, 624–638.

[20] BELLEC, P.C. and TSYBAKOV, A.B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.* 16, 1879–1892.

[21] BELLEC, P.C. (2018). Sharp oracle inequalities for Least Squares estimators in shape restricted regression. *Ann. Statist.* 46, 745–780.

[22] BHATTACHARYA, A., DANSON D.B., PATI D. and PILLAI, N.S. (2016). Sub-optimality of some continuous shrinkage priors. *Stoch. Proc. and their Appl.* 126, 3828–3842.

[23] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* 3, 203–268.

[24] BOUCHERON, S., LUGOSI, G., and MASSART, P. (2012). Concentration Inequalities (A nonasymptotic theory of independence). *Oxford University Press*

[25] BULL, A. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Statist.* 6, 1490–1516.

[26] BULL, A. and NICKL, R. (2013). Adaptive confidence sets in ℓ_2. *Probab. Theory and Rel. Fields.* 156, 889–919.

[27] BUNEA, F., TSYBAKOV, A.B., and WEGKAMP, M. (2007). Aggregation for Gaussian regression. *Ann. Stat.* 35, 1674–1697.

[28] CAI, T.T. and LOW, M.G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* 32, 1805–1840.

[29] CAI, T.T., ZHANG, C.-H. and ZHOU, H.H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* 38, 2118–2144.

[30] CAI, T.T. and ZHOU, H.H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* 40, 2389–2420.

[31] CARVALHO, C.M., POLSON N.G. and SCOTT J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465–480.

[32] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A.W. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* 43, 1986–2018.

[33] CASTILLO, I. and VAN DER VAART, A.W. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.* 40, 2069–2101.

[34] CAVALIER, L. and TSYBAKOV, A. (2001). Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Meth. Statist.* 10, 247–282.

[35] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* 43, 1774–1800.

[36] DONOHO, D.L., JOHNSTONE, I.M., HOCH, J.C. and STERN, A.S. (1992). Maximum entropy and the nearly black object(with Discussion). *J. Roy. Statist. Soc. Ser. B* 54, 41–81.

[37] DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–455.

[38] DONOHO, D.L. and JOHNSTONE, I.M. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields.* 99, 277–303.

[39] EFROMOVICH, S. and PINSKER, M. (1984). Learning algorithm for nonparametric filtering. *Automat. and Remote Control.* 24, 1434–1440.

[40] GAO, C., LU, Y. and ZHOU, H.H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* 43, 2624–2652.

[41] GAO, C., LU, Y., MA, Z. and ZHOU, H.H. (2016). Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.* 17, 1–29.

[42] GAO, C., VAN DER VAART, A.W. and ZHOU, H.H. (2015). A general framework for bayes structured linear models. ArXiv:1506.02174.

[43] GHOSAL, S., GHOSH, J. and VAN DER VAART, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* 28, 500–531.

[44] GHOSAL, S. and VAN DER VAART, A.W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* 29, 1233–1263.

[45] GHOSAL, S. and VAN DER VAART, A.W. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* 35, 192–223.

[46] HAN, Q. (2017). Bayes model selection. ArXiv:1704.07513.

[47] HARTIGAN, J.A. (1972). Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, 67, 123–129.

[48] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* 43, 2259–2295.

[49] JOHNSTONE, I.M. (2017). Gaussian estimation: Sequence and wavelet models. Book draft.

[50] JOHNSTONE, I. and SILVERMAN, B. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* 32, 1594–1649.

[51] KIRICHENKO, A. and VAN ZANTEN, J.H. (2017). Estimating a smooth function on a large graph by Bayesian Laplacian regularization. *Electron. J. Stat.* 11, 891–915.

[52] KIRICHENKO, A. and VAN ZANTEN, J.H. (2018). Minimax lower bounds for function estimation on graphs. *Electron. J. Stat.* 12, 651–666.

[53] KLOPP, O. (2015). Matrix completion by singular value thresholding: Sharp bounds. *Electron. J. Stat.* 9, 2348–2369.

[54] KLOPP, O., TSYBAKOV, A.B. and VERZELEN, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* 45, 316–354.

[55] KLOPP, O., LU, Y., TSYBAKOV, A.B. and ZHOU, H.H. (2017). Structured Matrix Estimation and Completion. ArXiv:1707.02090.

[56] KNEIP, A. (1994). Oracle linear smoothers. *Ann. Statist.* 22, 835–866.

[57] KOLAR, M. and LIU, H. (2012). Supplement to *Marginal regression for multitask learning. Proceedings of Machine Learning Research* 22, 647–655.

[58] LEPSKI, O. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* 35, 454–466.

[59] LEPSKI, O. and SPOKOINY, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* 6, 2512–2546.

[60] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* 17, 1001–1008.

[61] LOUNICI, K.,PONTIL, M., TSYBAKOV, A.B. and VAN DE GEER, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* 39, 2164–2204.

[62] MA, Z. and WU, Y. (2015). Volume Ratio, Sparsity, and Minimaxity Under Unitarily Invariant Norms. *IEEE Trans. Inform. Theory* 61, 6939–6956.

[63] MARTIN, R., MESS, R. and WALKER, S.G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* 23, 1822–1847.

[64] MARTIN, R. and WALKER, S.G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Statist.* 8, 2188–2206.

[65] MASSART, P. (2007). Concentration Inequalities and Model Selection. *Springer Lecture Notes in Mathematics.*

[66] NEMIROVSKI, A. (2000). Topics in Non-parametric Statistics *Springer, NY.*

[67] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* 41, 2852–2876.

[68] OLHEDE, S.C. and WOLFE, P.J. (2013). Nonparametric graphon estimation. ArXiv:1309.5936.

[69] OLHEDE, S.C. and WOLFE, P.J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* 111, 14722–14727.

[70] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* 28, 298–335.

[71] PINSKER, M. (1980). Optimal filtration of square-integrable signal in Gaussian white noise. *Problems Inform. Transmission.* 16, 120–133.

[72] PRESS, S.J. (1982). Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference. *Krieger Publishing Company.*

[73] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* 57, 6976–6994.

[74] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H.H. (2015) Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* 43, 991–1026.

[75] RIGOLLET, P. and TSYBAKOV, A.B. (2011). Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.* 39, 731–771.

[76] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis* 7, 311–334.

[77] ROBINS, J. and VAN DER VAART, A.W. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* 34, 229–253.

[78] ROČKOVÁ, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* 46, 401–437.

[79] ROUSSEAU, J. and SZABÓ, B. (2016). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. ArXiv:1609.05067.

[80] ROUSSEAU, J. and SZABÓ, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* 45, 833–865.

[81] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* 29, 687-714.

[82] SZABÓ, B. T., VAN DER VAART, A.W. and VAN ZANTEN, J.H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Statist.* 7, 991–1018.

[83] SZABÓ, B. T., VAN DER VAART, A.W. and VAN ZANTEN, J.H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* 43, 1391–1428.

[84] TSYBAKOV, A.B. (2014). Aggregation and minimax optimality in high-dimensional estimation. *Proceedings of the International Congress of Mathematicians*.

[85] VAN DER PAS, S.L., KLEIJN, B.J.K. and VAN DER VAART, A.W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* 8, 2585–2618.

[86] VAN DER PAS, S.L., SZABÓ, B. T. and VAN DER VAART, A.W. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* 12, 1221–1274.

[87] VAN DER VAART, A.W. and VAN ZANTEN, J.H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* 36, 1435–1463.